

Retropropagazione e apprendimento biologico

Parte I — Fondamenti matematici della retropropagazione dell'errore

Il problema dell'attribuzione del credito: storia e formulazione

La difficoltà di attribuire la responsabilità di un esito complessivo agli elementi intermedi di una catena causale è un problema che precede la neuroscienze computazionale di almeno un secolo. Gli economisti classici la riconoscevano nella questione del valore marginale dei fattori produttivi; i filosofi morali la incontravano nell'analisi dell'azione collettiva; i teorici del controllo la affrontavano nella progettazione di regolatori retroazionati. Nel contesto dell'apprendimento automatico, essa assunse per la prima volta una formulazione precisa con il lavoro di Rosenblatt sul percettrone (1958) e la successiva analisi di Minsky e Papert nel volume *Perceptrons* (1969), che dimostrava con rigore i limiti delle architetture a strato singolo proprio in ragione dell'impossibilità di attribuire correttamente il credito agli strati nascosti.

Il nucleo del problema è il seguente. Si consideri un sistema composto da N componenti $f_1, f_2, \dots, f$ disposti in serie, tale che l'uscita globale sia $y = f(f^{-1}(\dots f_1(x) \dots))$. Si osserva che y differisce da un valore desiderato y^* , e si vuole modificare i parametri interni di ciascun f_i in modo da ridurre questa discrepanza. La difficoltà è che l'errore è osservabile soltanto all'uscita del sistema, mentre i componenti interni producono soltanto stati intermedi mai direttamente confrontati con alcun obiettivo. Occorre dunque *risalire a ritroso* la catena causale per determinare in quale misura ciascun componente abbia contribuito all'errore finale.

Nel caso in cui tutte le f_i siano differenziabili, la soluzione è formalmente offerta dalla regola della catena del calcolo differenziale. Ma la *scoperta* che questa soluzione potesse essere calcolata in modo efficiente mediante un algoritmo lineare nel numero dei parametri — anziché richiedere una stima separata per ciascun parametro mediante perturbazioni finite — fu il contributo essenziale di Werbos (1974), Rumelhart, Hinton e Williams (1986) e Parker (1985). Werbos in particolare, nella sua dissertazione di dottorato ad Harvard, formulò l'algoritmo in termini di teoria del controllo ottimo, riconoscendone la connessione con il principio del massimo di Pontryagin e con la programmazione dinamica di Bellman. Questo collegamento con la teoria del controllo è tutt'altro che superficiale: la backpropagation è, dal punto di vista strutturale, un caso speciale del metodo delle equazioni aggiunte (*adjoint method*) usato nell'ottimizzazione di sistemi dinamici descritti da equazioni differenziali.

Grafi computazionali e differenziazione automatica

Prima di trattare la backpropagation nelle reti neurali stratificate, è istruttivo inquadrarla nel contesto più generale della *differenziazione automatica* (*automatic differentiation*, AD), di cui essa è un caso speciale. La differenziazione automatica è la disciplina che studia il calcolo esatto e sistematico delle derivate di funzioni espresse come programmi informatici.

Qualunque programma che calcoli una funzione differenziabile può essere rappresentato come un *grafo computazionale diretto aciclico* (*Directed Acyclic Graph*, DAG), in cui i nodi rappresentano operazioni elementari (somma, prodotto, applicazione di una funzione non lineare) e gli archi rappresentano il flusso dei valori intermedi. Formalmente, sia $v_1, v_2, \dots, v$ la sequenza topologicamente ordinata dei nodi del grafo, con $v_1, \dots, v$ che denotano le variabili di input e v la variabile di output (la funzione scalare da differenziare). Ogni nodo v_i con $i > k$ è definito da un'operazione elementare applicata a un sottoinsieme dei suoi predecessori:

$$v_i = \varphi_i(v : j \in \text{pa}(i))$$

dove $\text{pa}(i)$ denota l'insieme dei nodi genitori di i nel grafo.

La differenziazione automatica in modalità *reverse* — che è esattamente la backpropagation — calcola il gradiente $\partial v / \partial v_i$ per tutti i nodi i in un unico passaggio all'indietro sul grafo, sfruttando la seguente ricorrenza sulle *barre* (variabili aggiunte):

$$\bar{v}_i = \partial v / \partial v_i = \sum_{j \in \text{pa}(i)} \bar{v}_j \cdot \partial v / \partial v_j$$

L'inizializzazione è $\bar{v} = 1$. Questa ricorrenza accumula i contributi di tutti i cammini che dalla variabile v_i conducono all'uscita v , ponderati per le derivate parziali lungo ciascun arco. La complessità computazionale di questo passaggio è $O(|E|)$, dove $|E|$ è il numero di archi del grafo — una complessità che non dipende dal numero di variabili di input, a differenza della modalità *forward* che richiederebbe un passaggio per ciascuna variabile di input.

Questa proprietà è cruciale: per funzioni con molti input (come le reti neurali con milioni di parametri) e una sola uscita scalare (la funzione di perdita), la modalità *reverse* è di gran lunga più efficiente. Il costo di un singolo passaggio backward è proporzionale al costo del passaggio forward, tipicamente con una costante moltiplicativa di 2-3. Tutte le librerie di deep learning moderne — TensorFlow, PyTorch, JAX — implementano la backpropagation come un caso speciale di differenziazione automatica in modalità *reverse* su grafi computazionali.

Architettura formale: la rete neurale a propagazione in avanti

Specializziamo ora il framework generale al caso di una rete neurale artificiale con L strati nascosti più lo strato di uscita, per un totale di L strati parametrizzati (indicizzati $\ell = 1, 2, \dots, L$). Lo strato $\ell = 0$ è lo strato di input. Allo strato ℓ sono associate le seguenti strutture:

- una matrice di pesi $\mathbf{W}^\ell \in \mathbb{R}^{(n^\ell \times n^{\ell-1})}$, dove n^ℓ è il numero di unità nello strato ℓ ;
- un vettore di bias $\mathbf{b}^\ell \in \mathbb{R}^{(n^\ell)}$; — una funzione di attivazione $\sigma^\ell : \mathbb{R} \rightarrow \mathbb{R}$, applicata elemento per elemento.

L'insieme totale dei parametri è $\theta = \{\mathbf{W}^\ell, \mathbf{b}^\ell\}_{\ell=1}^L$, con dimensionalità $|\theta| = \sum_{\ell=1}^L (n^\ell \cdot n^{\ell-1} + n^\ell)$.

Il passaggio in avanti (*forward pass*) è definito dalla ricorrenza:

$$\begin{aligned} \mathbf{z}^{\ell+1} &= \mathbf{W}^{\ell+1} \mathbf{a}^{\ell+1-1} + \mathbf{b}^{\ell+1} && \in \mathbb{R}^{(n^{\ell+1})} \\ \mathbf{a}^{\ell+1} &= \sigma^{\ell+1}(\mathbf{z}^{\ell+1}) && \in \mathbb{R}^{(n^{\ell+1})} \end{aligned}$$

con $\mathbf{a}^0 = \mathbf{x} \in \mathbb{R}^{(n^0)}$. La quantità $\mathbf{z}^\ell$ è la *pre-attivazione* e $\mathbf{a}^\ell$ è l'*attivazione* dello strato ℓ . L'uscita della rete è $\hat{\mathbf{y}} = \mathbf{a}^L$.

È importante notare che questa struttura impone un'ipotesi di *separabilità*: la non-linearità σ^ℓ agisce separatamente su ciascuna coordinata di $\mathbf{z}^\ell$, e la parte lineare $\mathbf{W}^{\ell+1} \mathbf{a}^{\ell+1-1} + \mathbf{b}^{\ell+1}$ mischia le coordinate. Questa separazione tra mixing lineare e non-linearità coordinata è caratteristica delle reti dense; le reti convoluzionali, ricorrenti e transformer violano in modi diversi questa struttura, come vedremo.

Funzioni di perdita: geometria e proprietà analitiche

La funzione di perdita $\Delta : \mathbb{R}^{(n^L)} \times \mathbb{R}^{(n^L)} \rightarrow \mathbb{R}$ misura la discrepanza tra l'uscita $\hat{\mathbf{y}}$ e il target $\mathbf{y}$. La scelta è dettata sia da considerazioni statistiche (interpretazione probabilistica del modello) sia da considerazioni di ottimizzazione (convessità, comportamento del gradiente).

Errore quadratico medio (MSE). Per problemi di regressione:

$$\mathcal{L}_{\text{MSE}}(\hat{\mathbf{y}}, \mathbf{y}) = (1/2) \sum_i (\hat{y}_i - y_i)^2 = (1/2) \|\hat{\mathbf{y}} - \mathbf{y}\|^2$$

La divisione per 2 è convenzionale e semplifica la derivata. Il gradiente rispetto a $\hat{\mathbf{y}}$ è semplicemente:

$$\partial \mathcal{L}_{\text{MSE}} / \partial \hat{\mathbf{y}} = \hat{\mathbf{y}} - \mathbf{y}$$

Cross-entropia binaria. Per classificazione binaria con uscita $\sigma(z) \in (0,1)$:

$$\mathcal{L}_{\text{BCE}} = -[y \log \hat{y} + (1-y) \log(1-\hat{y})]$$

Il gradiente è:

$$\partial \mathcal{L}_{\text{BCE}} / \partial \hat{y} = -y/\hat{y} + (1-y)/(1-\hat{y}) = (\hat{y} - y)/(\hat{y}(1-\hat{y}))$$

Quando si combina con la sigmoide $\sigma(z) = 1/(1+e^{-z})$ come funzione di attivazione dell'ultimo strato, il prodotto $\partial \mathcal{L} / \partial \hat{y} \cdot \sigma'(z)$ si semplifica notevolmente:

$$\partial \mathcal{L}_{\text{BCE}} / \partial z = [\partial \mathcal{L}_{\text{BCE}} / \partial \hat{y}] \cdot [\hat{y}(1-\hat{y})] = \hat{y} - y$$

una cancellazione algebrica che elimina il termine di denominatore e produce un gradiente numericamente stabile.

Cross-entropia categoriale con softmax. Per classificazione multiclasse con C classi, la funzione softmax è:

$$\hat{y}_i = \exp(z_i) / \sum \exp(z) \quad \text{per } i = 1, \dots, C$$

e la perdita è:

$$\mathcal{L}_{\text{CE}} = -\sum_i y_i \log \hat{y}_i$$

dove $\mathbf{y}$ è un vettore one-hot. Il gradiente rispetto a $\mathbf{z}$ richiede il calcolo del Jacobiano della softmax. Per la softmax, le derivate parziali sono:

$$\partial \hat{y}_i / \partial z = \hat{y}_i(\delta_{ii} - \hat{y})$$

dove δ_{ii} è il delta di Kronecker. Quindi:

$$\begin{aligned} \partial \mathcal{L}_{\text{CE}} / \partial z &= \sum_i (\partial \mathcal{L}_{\text{CE}} / \partial \hat{y}_i) (\partial \hat{y}_i / \partial z) \\ &= \sum_i (-y_i / \hat{y}_i) \cdot \hat{y}_i(\delta_{ii} - \hat{y}) \\ &= \sum_i (-y_i)(\delta_{ii} - \hat{y}) \\ &= -y + \hat{y} \sum_i y_i \\ &= \hat{y} - y \end{aligned}$$

dove si è usato il fatto che $\sum_i y_i = 1$ per un target one-hot. In forma vettoriale:

$$\partial \mathcal{L}_{\text{CE}} / \partial \mathbf{z} = \hat{\mathbf{y}} - \mathbf{y}$$

Questa formula, identica a quella dell'MSE ma derivante da premesse completamente diverse, è uno degli esempi più eleganti di cancellazione algebrica in tutta la teoria dell'ottimizzazione delle reti neurali. Essa garantisce che il segnale di gradiente all'uscita sia proporzionale

all'errore di classificazione, senza saturazione — a differenza di quanto accadrebbe usando MSE con un'uscita softmax.

Perdita di Huber. Per regressione robusta agli outlier, la perdita di Huber interpola tra MSE e MAE:

$$\begin{aligned}\mathcal{L}_{\text{Huber}}(r) &= (1/2)r^2 && \text{se } |r| \leq \delta \\ \mathcal{L}_{\text{Huber}}(r) &= \delta|r| - (1/2)\delta^2 && \text{se } |r| > \delta\end{aligned}$$

dove $r = \hat{y} - y$. Il gradiente è continuo in $r = \pm\delta$ ma non differenziabile in quei punti:

$$\begin{aligned}\partial \mathcal{L}_{\text{Huber}} / \partial r &= r && \text{se } |r| \leq \delta \\ \partial \mathcal{L}_{\text{Huber}} / \partial r &= \delta \operatorname{sign}(r) && \text{se } |r| > \delta\end{aligned}$$

Derivazione completa dell'algoritmo: i segnali di errore δ^l

Il cardine algebrico della backpropagation è il vettore di segnale di errore δ^l , definito come il gradiente della perdita rispetto alla pre-attivazione dello strato ℓ :

$$\delta^{**l} \triangleq \partial \mathcal{L} / \partial z^{**l} \in \mathbb{R}^{(n^l)}$$

Questa scelta di variabile ausiliaria non è arbitraria: è quella che permette di esprimere i gradienti rispetto a tutti i parametri in forma più compatta, e che rivela la struttura ricorsiva del calcolo backward.

Caso base: strato di uscita $\ell = L$.

Dalla definizione e dalla regola della catena:

$$\delta_i^L = \partial \mathcal{L} / \partial z_i^L = \sum (\partial \mathcal{L} / \partial a_i^L) (\partial a_i^L / \partial z_i^L)$$

Poiché $a_i^L = \sigma^L(z_i^L)$ e σ^L agisce componente per componente, $\partial a_i^L / \partial z_i^L = \sigma'^L(z_i^L) \delta_{ii}$ (delta di Kronecker). Quindi:

$$\delta_i^L = (\partial \mathcal{L} / \partial a_i^L) \cdot \sigma'^L(z_i^L)$$

In forma vettoriale:

$$\delta^{**L} = (\nabla_{a^{**L}} \mathcal{L}) \odot \sigma'^L(z^{**L})$$

dove $\odot$ è il prodotto di Hadamard (elemento per elemento) e $\sigma^l(\mathbf{z}^l)$ denota il vettore delle derivate componente per componente.

Passo ricorsivo: strato $\ell < L$.

Occorre esprimere δ^ℓ in termini di $\delta^{\ell+1}$. La connessione tra i due strati è:

$$\mathbf{z}^{\ell+1} = \mathbf{W}^{\ell+1} \mathbf{a}^\ell + \mathbf{b}^{\ell+1} = \mathbf{W}^{\ell+1} \sigma^l(\mathbf{z}^\ell) + \mathbf{b}^{\ell+1}$$

Quindi:

$$\begin{aligned} \partial \mathcal{L} / \partial z_i^\ell &= \sum (\partial \mathcal{L} / \partial z_i^{\ell+1}) (\partial z_i^{\ell+1} / \partial z_i^\ell) \\ &= \sum \delta_i^{\ell+1} \cdot W_{i,j}^{\ell+1} \cdot \sigma_j^l(z_i^\ell) \\ &= \sigma_j^l(z_i^\ell) \cdot \sum W_{i,j}^{\ell+1} \cdot \delta_j^{\ell+1} \end{aligned}$$

In forma matriciale, riconoscendo che $\sum W_{i,j}^{\ell+1} \cdot \delta_j^{\ell+1}$ è la i -esima componente del prodotto $(\mathbf{W}^{\ell+1})^T \delta^{\ell+1}$:

$$\delta^\ell = ((\mathbf{W}^{\ell+1})^T \delta^{\ell+1}) \odot \sigma^l(\mathbf{z}^\ell)$$

Questa è la *formula di ricorrenza della backpropagation*. Il suo significato geometrico è netto: il segnale di errore dallo strato superiore viene proiettato nello spazio delle pre-attivazioni dello strato inferiore mediante la trasposta della matrice di connessione, e poi modulato puntualmente dalla derivata della non-linearità locale.

Gradienti rispetto ai parametri.

Per i pesi dello strato ℓ , partendo da $\mathbf{z}^\ell = \mathbf{W}^\ell \mathbf{a}^{\ell-1} + \mathbf{b}^\ell$:

$$\partial \mathcal{L} / \partial W_{i,j}^\ell = \sum (\partial \mathcal{L} / \partial z_i^\ell) (\partial z_i^\ell / \partial W_{i,j}^\ell) = \delta_i^\ell \cdot a_j^{\ell-1}$$

In forma matriciale, questa è una matrice ottenuta dal prodotto esterno:

$$\partial \mathcal{L} / \partial \mathbf{W}^\ell = \delta^\ell (\mathbf{a}^{\ell-1})^T \in \mathbb{R}^{(n^\ell \times n^{\ell-1})}$$

Per i bias:

$$\partial \mathcal{L} / \partial b_i^\ell = \sum (\partial \mathcal{L} / \partial z_i^\ell) (\partial z_i^\ell / \partial b_i^\ell) = \delta_i^\ell \cdot 1 = \delta_i^\ell$$

quindi:

$$\partial \mathcal{L} / \partial \mathbf{b}^\ell = \delta^\ell$$

Nel caso di mini-batch. Per un mini-batch di m esempi con input $\mathbf{X} \in \mathbb{R}^{(n^0 \times m)}$ e target $\mathbf{Y} \in \mathbb{R}^{(n^L \times m)}$, le attivazioni diventano matrici $\mathbf{A}^l \in \mathbb{R}^{(n^l \times m)}$ e le formule si generalizzano:

$$\mathbf{Z}^{**l} = \mathbf{W}^{**l} \mathbf{A}^{**l-1} + \mathbf{b}^{**l} \mathbf{1}^{**T} \quad (\text{broadcasting del bias})$$

$$\mathbf{A}^{**l} = \sigma^l(\mathbf{Z}^{**l})$$

$$\Delta^{**L} = (\partial \mathcal{L} / \partial \mathbf{A}^{**L}) \odot \sigma'^L(\mathbf{Z}^{**L})$$

$$\Delta^{**l} = ((\mathbf{W}^{**l+1})^T \Delta^{**l+1}) \odot \sigma'^l(\mathbf{Z}^{**l})$$

$$\partial \mathcal{L} / \partial \mathbf{W}^{**l} = (1/m) \Delta^{**l} (\mathbf{A}^{**l-1})^T$$

$$\partial \mathcal{L} / \partial \mathbf{b}^{**l} = (1/m) \Delta^{**l} \mathbf{1}^{**}$$

dove $\mathbf{1} \in \mathbb{R}^m$ è il vettore di tutti uno e la media $1/m$ normalizza per la dimensione del batch.

Funzioni di attivazione: analisi approfondita

La funzione di attivazione è il componente non lineare che conferisce alle reti neurali la loro capacità espressiva universale. Senza non-linearità, qualunque rete profonda ridurrebbe a una trasformazione lineare, indipendentemente dal numero di strati. Il teorema di approssimazione universale (Cybenko, 1989; Hornik, 1991) garantisce che reti con almeno uno strato nascosto dotato di attivazioni sigmoidei possono approssimare uniformemente qualunque funzione continua su un compatto, con accuratezza arbitraria al crescere della larghezza. Questo risultato, sebbene di natura esistenziale (non costruttiva), ha sancito la legittimità teorica delle reti neurali come approssimatori universali.

Sigmoide logistica. $\sigma(z) = (1 + e^{-z})^{-1}$

Range: $(0, 1)$. Proprietà importanti:

$$\sigma'(z) = \sigma(z)(1 - \sigma(z))$$

$$\sigma(-z) = 1 - \sigma(z) \quad (\text{antisimmetria rispetto al punto } (0, 1/2))$$

$$\sigma''(z) = \sigma'(z)(1 - 2\sigma(z))$$

$$\int_{-\infty}^{\infty} \sigma'(z) dz = 1$$

Il massimo di $\sigma'(z)$ è $1/4$, in $z = 0$. Per $|z| > 3$, $\sigma'(z) < 0.045$, rendendo il gradiente trascurabile. Questa saturazione bilaterale produce il problema del vanishing gradient nelle reti profonde con attivazioni sigmoidei.

Tangente iperbolica. $\sigma(z) = \tanh(z) = (e^z - e^{-z}) / (e^z + e^{-z})$

Relazione con la sigmoide: $\tanh(z) = 2\sigma(2z) - 1$. Range: $(-1, 1)$. Derivata:

$$\tanh'(z) = 1 - \tanh^2(z) = \text{sech}^2(z)$$

Il massimo di $\tanh'(z)$ è 1 in $z = 0$. La simmetria attorno all'origine — $\tanh(-z) = -\tanh(z)$ — riduce il problema delle correlazioni tra gradienti di strati successivi, rendendo $\tanh$ preferibile alla sigmoide nelle applicazioni pratiche prima dell'avvento delle ReLU.

ReLU e varianti. Rectified Linear Unit: $\sigma(z) = \max(0, z)$.

$$\sigma'(z) = \mathbb{1}[z > 0] \quad (\text{subgradiente: } 0 \text{ in } z = 0 \text{ per convenzione})$$

La ReLU è computazionalmente efficiente, produce gradienti non saturi per $z > 0$, e induce sparsità naturale nelle attivazioni (circa il 50% delle unità è silenziata per input casuali gaussiani). La sua principale debolezza è il *dying ReLU*: se una unità accumula un bias negativo sufficiente, $z < 0$ per tutti gli input di training, $\sigma'(z) = 0$ costantemente, e nessun aggiornamento di gradiente raggiungerà le sinapsi a monte.

Leaky ReLU: $\sigma(z) = \max(\alpha z, z)$ con $\alpha \in (0, 1)$, tipicamente $\alpha = 0.01$.

$$\sigma'(z) = 1 \text{ se } z \geq 0, \alpha \text{ se } z < 0$$

Parametric ReLU (PReLU): come Leaky ReLU ma con α appreso per retropropagazione.

ELU (Exponential Linear Unit): $\sigma(z) = z$ se $z \geq 0$, $\alpha(e^z - 1)$ se $z < 0$.

$$\sigma'(z) = 1 \text{ se } z \geq 0, \alpha e^z \text{ se } z < 0$$

ELU ha un output medio più vicino a zero rispetto alla ReLU, riducendo il *bias shift* nelle attivazioni degli strati successivi, a costo di un calcolo esponenziale per $z < 0$.

SiLU (Sigmoid Linear Unit) / Swish: $\sigma(z) = z \cdot \sigma_{\text{logistic}}(z)$.

$$\begin{aligned} \sigma'(z) &= \sigma_{\text{logistic}}(z) + z \cdot \sigma_{\text{logistic}}(z)(1 - \sigma_{\text{logistic}}(z)) \\ &= \sigma_{\text{logistic}}(z)(1 + z(1 - \sigma_{\text{logistic}}(z))) \end{aligned}$$

La Swish è non monotona — mostra un minimo locale intorno a $z \approx -1.28$ — e differenziabile ovunque. Ramachandran et al. (2017) hanno dimostrato che supera empiricamente la ReLU su molti benchmark in modelli profondi.

GELU (Gaussian Error Linear Unit). Hendrycks e Gimpel (2016):

$$\sigma_{\text{GELU}}(z) = z \cdot \Phi(z) = z \cdot (1/2)[1 + \text{erf}(z/\sqrt{2})]$$

dove Φ è la CDF della gaussiana standard ed erf è la funzione errore. Approssimazione computazionalmente efficiente:

$$\sigma_{\text{GELU}}(z) \approx 0.5z(1 + \tanh[\sqrt{2/\pi}(z + 0.044715z^3)])$$

La derivata è:

$$\sigma'_{\text{GELU}}(z) = \Phi(z) + z \cdot \varphi(z)$$

dove φ è la PDF gaussiana standard. La GELU può essere interpretata come un'aspettativa stocastica: corrisponde ad applicare una ReLU a z perturbato da rumore gaussiano, e in questa luce implementa implicitamente il dropout durante il forward pass. È la funzione di attivazione predominante nei modelli di linguaggio di grandi dimensioni basati su transformer (BERT, GPT, T5).

Analisi del vanishing e exploding gradient

Il problema del gradiente che svanisce o esplode è il principale ostacolo all'addestramento di reti profonde, e merita un trattamento analitico rigoroso che vada oltre la semplice osservazione qualitativa.

Modello semplificato. Consideriamo una rete con L strati, tutti con la stessa funzione di attivazione σ e gli stessi pesi scalari w (caso unidimensionale per semplicità). Le equazioni di forward e backward sono:

$$\begin{aligned} z^l &= w \cdot a^{l-1} + b, \quad a^l = \sigma(z^l) \\ \delta^l &= \delta^{l+1} \cdot w \cdot \sigma'(z^l) \end{aligned}$$

Quindi:

$$\delta^1 = \delta^L \cdot \prod_{l=2}^L [w \cdot \sigma'(z^l)] = \delta^L \cdot w^{L-1} \cdot \prod_{l=2}^L \sigma'(z^l)$$

Se $\sigma = \tanh$ e le pre-attivazioni z^l sono piccole (regime lineare), $\sigma'(z^l) \approx 1$. Allora $\delta^1 \approx \delta^L \cdot w^{L-1}$. Se $|w| < 1$, il gradiente decade esponenzialmente con L ; se $|w| > 1$, cresce esponenzialmente.

Caso multidimensionale: raggio spettrale. Per reti con n^l unità per strato, il fattore di scala per ogni strato è la matrice $\mathbf{W}^{l+1} \cdot \text{diag}(\sigma'(z^l))$. Il comportamento del gradiente è governato dal raggio spettrale $\varrho = \varrho(\mathbf{W}^l \cdot \text{diag}(\sigma'))$:

— se $\varrho < 1$ per ogni strato, il gradiente converge a zero esponenzialmente; — se $\varrho > 1$ per ogni strato, il gradiente diverge esponenzialmente; — solo se $\varrho \approx 1$ per ogni strato (regime critico), il gradiente rimane stabile.

Mantenere il raggio spettrale vicino a 1 durante l'addestramento è tecnicamente difficile, perché i pesi cambiano a ogni iterazione. Le due principali strategie per mitigare questo problema sono l'inizializzazione opportuna e la normalizzazione degli strati.

Inizializzazione di Xavier/Glorot. Glorot e Bengio (2010) derivano la varianza ottimale per l'inizializzazione dei pesi richiedendo che la varianza delle attivazioni e dei gradienti rimanga costante attraverso gli strati. Per attivazioni con $\sigma'(0) = 1$ (come tanh nel regime lineare), la condizione di equilibrio della varianza delle attivazioni impone:

$$\text{Var}[a^l] = \text{Var}[a^{l-1}] \Rightarrow n^{l-1} \cdot \text{Var}[\mathbf{W}^l] = 1 \Rightarrow \text{Var}[\mathbf{W}^l] = 1/n^{l-1}$$

Analogamente, la condizione di equilibrio della varianza dei gradienti impone:

$$\text{Var}[\mathbf{W}^l] = 1/n^l$$

Il compromesso suggerito da Glorot è:

$$\text{Var}[\mathbf{W}^l] = 2/(n^{l-1} + n^l)$$

che bilancia i due requisiti. I pesi sono quindi campionati da $U(-\sqrt{6/(n^{l-1}+n^l)}, +\sqrt{6/(n^{l-1}+n^l)})$ o equivalentemente da $\mathcal{N}(0, 2/(n^{l-1}+n^l))$.

Inizializzazione di He. Per le ReLU, $\sigma'(z) = \mathbb{1}[z>0] \approx 1/2$ in media, dimezzando la varianza a ogni strato. He et al. (2015) derivano:

$$\text{Var}[\mathbf{W}^l] = 2/n^{l-1}$$

da cui i pesi vengono campionati da $\mathcal{N}(0, 2/n^{l-1})$. Questa è l'inizializzazione standard per architetture con attivazioni ReLU o sue varianti.

Batch Normalization. Ioffe e Szegedy (2015) propongono di normalizzare esplicitamente le pre-attivazioni a ogni strato prima di applicare la non-linearità. Per un mini-batch $\{\mathbf{z}^{l(i)}\}_{i=1}^m$, la trasformazione è:

$$\begin{aligned}\mu^l &= (1/m) \sum_i \mathbf{z}^{l(i)} \\ \sigma^{2l} &= (1/m) \sum_i (\mathbf{z}^{l(i)} - \mu^l)^2 \\ \hat{\mathbf{z}}^{l(i)} &= (\mathbf{z}^{l(i)} - \mu^l) / \sqrt{\sigma^{2l} + \epsilon} \\ \tilde{\mathbf{a}}^{l(i)} &= \gamma^l \odot \hat{\mathbf{z}}^{l(i)} + \beta^l\end{aligned}$$

dove γ^l e β^l sono parametri appresi (scala e traslazione), e $\epsilon > 0$ è una costante per la stabilità numerica. La BN stabilizza la distribuzione delle attivazioni e mantiene il raggio spettrale effettivo vicino a 1, eliminando di fatto il vanishing gradient nelle architetture profonde. I

gradienti rispetto a γ^l , β^l , μ^l , σ^{2l} vengono anch'essi calcolati per backpropagation, ma richiedono attenzione per la dipendenza tra esempi nel batch.

Backpropagation through time: derivazione completa

Per le reti neurali ricorrenti, il grafo computazionale ha una struttura ciclica che viene resa aciclica dispiegando la rete nel tempo. Consideriamo la RNN più semplice:

$$\begin{aligned} \mathbf{h}^{**} &= \sigma(\mathbf{W}_{hh}^{**} \mathbf{h}_{-1}^{**} + \mathbf{W}_{xh}^{**} \mathbf{x}^{**} + \mathbf{b}_{hh}^{**}) \\ \mathbf{o}^{**} &= \mathbf{W}_{hy}^{**} \mathbf{h}^{**} + \mathbf{b}_{hy}^{**} \\ \hat{\mathbf{y}}^{**} &= \text{softmax}(\mathbf{o}^{**}) \end{aligned}$$

La perdita totale per una sequenza di T passi è:

$$\mathcal{L} = \sum_{t=1}^T \mathcal{L}(\hat{\mathbf{y}}^{**}, \mathbf{y}^{**})$$

Gradiente rispetto a $\mathbf{W}_{hy}$ e $\mathbf{b}_{hy}$. Poiché $\mathbf{o}$ dipende da $\mathbf{W}_{hy}$ soltanto attraverso $\mathbf{h}$, e i diversi timestep condividono gli stessi parametri:

$$\partial \mathcal{L} / \partial \mathbf{W}_{hy}^{**} = \sum_{t=1}^T (\partial \mathcal{L} / \partial \mathbf{o}^{**}) \otimes \mathbf{h}^{**} = \sum_{t=1}^T \delta^{**} \mathbf{h}^{**T}$$

Gradiente rispetto a $\mathbf{W}_{hh}$ e $\mathbf{W}_{xh}$. Questi parametri influenzano $\mathcal{L}$ attraverso tutti gli stati nascosti $\mathbf{h}_s$ con $s \leq t$. Il gradiente dell'errore rispetto a $\mathbf{h}$ è:

$$\partial \mathcal{L} / \partial \mathbf{h}^{**}_t = \partial \mathcal{L} / \partial \mathbf{h}^{**}_t + (\partial \mathcal{L} / \partial \mathbf{h}^{**}_{t+1}) (\partial \mathbf{h}^{**}_{t+1} / \partial \mathbf{h}^{**}_t)$$

ossia, in forma ricorrente all'indietro (da $t = T$ a $t = 1$):

$$\delta^{**}_t \mathbf{h} = \mathbf{W}_{hy}^{**T} \delta^{**}_o + \mathbf{W}_{hh}^{**T} \text{diag}(\sigma'(\mathbf{z}^{**}_{t+1})) \delta^{**}_{t+1} \mathbf{h}$$

con $\delta_t^h = \partial \mathcal{L} / \partial \mathbf{h}$ e $\delta_o = \partial \mathcal{L} / \partial \mathbf{o}$. Il contributo al gradiente di $\mathbf{W}_{hh}$ da ogni timestep t è:

$$\partial \mathcal{L} / \partial \mathbf{W}_{hh}^{**} = \sum_{t=1}^T (\partial \mathcal{L} / \partial \mathbf{h}^{**}_t) \cdot (\prod_{s=t+1}^T \text{diag}(\sigma'(\mathbf{z}^{**}_s)) \mathbf{W}_{hh}^{**}) \cdot \text{diag}(\sigma'(\mathbf{z}^{**}_t)) \cdot \mathbf{h}^{**}_{t-1}^T$$

Il prodotto $\prod_{s=t+1}^T \text{diag}(\sigma'(\mathbf{z}^{**}_s)) \mathbf{W}_{hh}^{**}$ è un prodotto di $t-s$ matrici, soggetto ai già discussi problemi di svanimento ed esplosione. Per sequenze lunghe ($T > 100$), il gradiente delle perdite temporali lontane rispetto ai parametri ricorrenti diventa computazionalmente irrecuperabile con attivazioni sigmoidi o tanh.

Gradient clipping. Una tecnica empirica efficace contro il gradiente esplosivo nelle RNN è il *gradient clipping* di Pascanu et al. (2013): si calcola la norma del gradiente totale $\|g\|$ e, se supera una soglia θ , lo si riscalda:

$$g \leftarrow g \cdot \theta / \|g\| \quad \text{se } \|g\| > \theta$$

Questo non elimina il problema del gradiente esplosivo ma ne limita l'effetto sugli aggiornamenti dei parametri, stabilizzando l'addestramento.

Reti convoluzionali: backpropagation con struttura locale

Le reti neurali convoluzionali (CNN) sono caratterizzate da connettività locale e condivisione dei pesi, che introducono vincoli strutturali nella matrice W^l . In una convoluzione discreta, il filtro $k \in \mathbb{R}^{(d \times d)}$ è applicato all'input $a^{l-1} \in \mathbb{R}^{(H \times W)}$ per produrre una feature map:

$$z^l[i,j] = \sum_m \sum_n k[m,n] \cdot a^{l-1}[i+m, j+n] + b$$

La backpropagation in questo contesto richiede il calcolo del gradiente rispetto al filtro k e rispetto all'input a^{l-1} .

Gradiente rispetto al filtro. Il gradiente rispetto all'elemento $k[m,n]$ del filtro è:

$$\partial \mathcal{L} / \partial k[m,n] = \sum_i \sum_j \delta^l[i,j] \cdot a^{l-1}[i+m, j+n]$$

che è una correlazione crociata tra il segnale di errore δ^l e l'input a^{l-1} — strutturalmente identica a una convoluzione, ma con un segnale diverso.

Gradiente rispetto all'input (per la propagazione backward). Il segnale di errore da propagare allo strato inferiore è:

$$\delta^{l-1}[i,j] = \sum_m \sum_n k[m,n] \cdot \delta^l[i-m, j-n]$$

che è una convoluzione *completa* (full convolution) del segnale di errore δ^l con il filtro k ruotato di 180°. In notazione simbolica:

$$g^{l-1} = k^{**}_{\text{rotated}} \star g^l$$

dove $\star$ denota la convoluzione completa. Questo risultato ha un'interpretazione intuitiva: la risposta di un neurone convoluzionale all'offset (m,n) del filtro influenza tutti i neuroni superiori nel cui campo recettivo rientra quell'offset; il gradiente backward accumula questi contributi mediante una convoluzione inversa.

Normalizzazione, ottimizzatori avanzati e geometria del gradiente

La discesa del gradiente nella sua forma più elementare — $\theta \leftarrow \theta - \eta \nabla_{\theta} \mathcal{L}$ — presuppone che la funzione di perdita abbia curvatura uniforme in tutte le direzioni, il che è raramente vero per le perdite delle reti neurali. La matrice hessiana $H = \partial^2 \mathcal{L} / \partial \theta^2$ ha tipicamente autovalori distribuiti su molti ordini di grandezza, rendendo la direzione del gradiente scarsamente allineata con la direzione di massima discesa *effettiva* rispetto alla geometria della superficie di perdita.

Il metodo di Newton e le sue approssimazioni. Il metodo di Newton corregge per la curvatura moltiplicando il gradiente per l'inversa dell'hessiana:

$$\theta^{**} \leftarrow \theta^{**} - \eta H^{*-1} \nabla_{\theta^{**}} \mathcal{L}$$

Questo aggiornamento è invariante a riscaldamento lineare dei parametri ed è molto più efficiente della discesa del gradiente per superfici altamente anisotrope. Tuttavia, per reti con milioni di parametri, il calcolo e l'inversione di $H \in \mathbb{R}^{(|\theta| \times |\theta|)}$ è computazionalmente proibitivo.

Metodi di primo ordine adattativi. I metodi moderni approssimano la curvatura usando soltanto informazione del primo ordine accumulata durante l'addestramento. AdaGrad (Duchi et al., 2011) accumula la somma dei quadrati dei gradienti passati:

$$G_i = \sum_{j=1}^i (\partial \mathcal{L} / \partial \theta_j)^2$$

$$\theta_i \leftarrow \theta_i - (\eta / \sqrt{G_i + \epsilon}) \cdot (\partial \mathcal{L} / \partial \theta_i)$$

RMSProp (Hinton, 2012) sostituisce la somma cumulativa con una media mobile esponenziale per evitare che il tasso di apprendimento scenda a zero:

$$E_i = \rho E_{i-1} + (1-\rho)(\partial \mathcal{L} / \partial \theta_i)^2$$

$$\theta_i \leftarrow \theta_i - (\eta / \sqrt{E_i + \epsilon}) \cdot (\partial \mathcal{L} / \partial \theta_i)$$

Adam (Kingma e Ba, 2015) combina media mobile dei gradienti (momento del primo ordine) e media mobile dei gradienti al quadrato (momento del secondo ordine), con correzione del bias per i passi iniziali:

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t \quad (\text{primo momento})$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2 \quad (\text{secondo momento})$$

$$\hat{m}_t = m_t / (1-\beta_1^t) \quad (\text{correzione del bias})$$

$$\hat{v}_t = v_t / (1-\beta_2^t) \quad (\text{correzione del bias})$$

$$\theta_{t+1} = \theta_t - \eta \hat{m}_t / (\sqrt{\hat{v}_t} + \epsilon)$$

I valori tipici dei parametri sono $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$. Adam approssima l'inversa della matrice di Fisher dell'informazione diagonalizzata, collegandosi alla *natural gradient descent* di Amari (1998), che ottimizza nella metrica dello spazio dei parametri definita dalla struttura probabilistica del modello.

La simmetria dei pesi e le sue implicazioni profonde

Ritorniamo sulla proprietà strutturale più critica per il confronto biologico. La formula di ricorrenza

$$^{**}\delta^{**l} = ((^{**}\mathbf{W}^{**l+1})^T ^{**}\delta^{**l+1}) \odot \sigma'(^{**}\mathbf{z}^{**l})$$

utilizza $(\mathbf{W}^{l+1})^T$ — la trasposta della matrice di pesi feedforward — per propagare il segnale di errore nella direzione opposta al flusso di informazione. In un grafo computazionale, questo corrisponde al fatto che il gradiente *percorre ogni arco nella direzione inversa* rispetto al forward pass, con un peso che è esattamente l'intensità della connessione nel forward pass.

Se si interpreta ogni peso $\mathbf{W}^{l+1}_{ij}$ come l'efficacia di una sinapsi eccitatoria dal neurone j dello strato l al neurone i dello strato $l+1$, la backpropagation richiede che esista una sinapsi *retrograda* con la stessa efficacia $\mathbf{W}^{l+1}_{ij}$ dal neurone i al neurone j . Questa sinapsi retrograda non deve essere un'entità fisica distinta — è sufficiente che il segnale di errore δ^{l+1}_i , giunto al neurone i dello strato $l+1$, venga trasmesso al neurone j dello strato l con un guadagno proporzionale a $\mathbf{W}^{l+1}_{ij}$.

Nella neurobiologia, le connessioni reciproche tra aree corticali esistono abbondantemente (le proiezioni feedback sono anzi quantitativamente superiori alle feedforward nella corteccia), ma la loro efficacia non è correlata in modo sistematico con quella delle connessioni corrispondenti nella direzione opposta. Le proiezioni feedforward originano prevalentemente dagli strati superficiali (II/III) e terminano sugli strati granulari (IV) dell'area ricevente; le proiezioni feedback originano dagli strati profondi (V/VI) e terminano sugli strati I e VI. Sono circuiti anatomicamente distinti, con caratteristiche di plasticità differenti, e non vi è alcuna ragione evolutiva o biofisica per cui le loro efficacie debbano essere vincolate a essere uguali.

Questa considerazione non è soltanto una critica alla plausibilità biologica della backpropagation: è anche la porta d'accesso verso una classe di algoritmi alternativi — la *feedback alignment* e i suoi discendenti — che esplorano se e quanto la simmetria dei pesi sia *necessaria* per l'apprendimento efficace, oppure soltanto *sufficiente*.

Parte II — L'implausibilità biologica e le teorie alternative: analisi approfondita

Fisiologia sinaptica: substrato per qualunque teoria dell'apprendimento

Prima di analizzare le obiezioni all'ipotesi biologica della backpropagation, è indispensabile avere un quadro preciso di come le sinapsi biologiche funzionano e cambiano. La sinapsi chimica è la giunzione tra l'assone di un neurone pre-sinaptico e il soma, i dendriti o i terminali assonici di un neurone post-sinaptico. La trasmissione sinaptica avviene in quattro fasi: depolarizzazione del terminale pre-sinaptico da un potenziale d'azione, apertura dei canali del calcio voltaggio-dipendenti, fusione delle vescicole contenenti neurotrasmettitori con la membrana pre-sinaptica (*exocytosis*) e legame del neurotrasmettitore rilasciato con i recettori post-sinaptici.

Le due principali classi di neurotrasmettitori nella corteccia sono il glutammato (eccitatorio) e il GABA (γ -aminobutirrico, inibitorio). I recettori ionotropici del glutammato più rilevanti per la plasticità sono i recettori AMPA e NMDA. I recettori AMPA sono permeabili ai cationi Na^+ e K^+ e mediano la risposta sinaptica rapida (correnti postsinaptiche eccitatorie, EPSC) su scala temporale di 1-5 ms. I recettori NMDA sono particolarmente significativi: sono sia voltaggio-dipendenti (bloccati da Mg^{2+} a potenziali di membrana di riposo) sia ligando-dipendenti (richiedono il legame del glutammato), il che li rende rivelatori di coincidenza (*coincidence detectors*) tra attività pre- e post-sinaptica. L'ingresso di Ca^{2+} attraverso i recettori NMDA attiva la CaM-kinasi II (CaMKII), una molecola fondamentale per l'espressione della LTP.

La potenza sinaptica — misurata come l'ampiezza del potenziale post-sinaptico eccitatorio (EPSP) prodotto da un singolo potenziale d'azione pre-sinaptico — è determinata da molteplici fattori: la probabilità di rilascio del neurotrasmettitore (che varia tra 0.1 e 0.9 a seconda del tipo di sinapsi), il numero e il tipo di recettori post-sinaptici attivati, la resistenza di input locale della membrana post-sinaptica, e la distanza dal soma. Questa molteplicità di variabili introduce una variabilità stocastica considerevole: il coefficiente di variazione della risposta a singolo spike può essere del 30-100% anche per sinapsi stabili. La modifica sinaptica durante l'apprendimento agisce su alcuni di questi fattori — aumentando la densità dei recettori AMPA post-sinaptici nella LTP, riducendola nella LTD — ma non li azzerava.

Le cinque obiezioni: trattamento analitico

Prima obiezione: il problema del trasporto dei pesi.

La matrice $(\mathbf{W}^{l+1})^T$ nella ricorrenza di backpropagation impone che il peso della connessione retrograda da j a i sia uguale al peso della connessione anterograda da i a j . Formalmente, se definiamo la matrice di connettività feedback come $\mathbf{B}^l \in \mathbb{R}^{(n^l \times n^{l+1})}$, la backpropagation richiede:

$$**\mathbf{B}^{*l} = (**\mathbf{W}^{*l+1})^T$$

Questa identità deve valere esattamente e deve essere mantenuta durante tutto l'addestramento, mentre i pesi $\mathbf{W}^{l+1}$ cambiano. Non è sufficiente che $\mathbf{B}^l$ sia *proporzionale* a $(\mathbf{W}^{l+1})^T$ o *allineata* con essa: deve esserne la trasposta esatta.

In un substrato sinaptico biologico, questa identità richiederebbe un meccanismo di accoppiamento tra la sinapsi anterograda (dal neurone A al neurone B) e la sinapsi retrograda (da B ad A) che: (a) garantisca la loro efficacia relativa in ogni momento, (b) aggiorni automaticamente l'efficacia della connessione retrograda ogni volta che quella anterograda viene modificata dall'apprendimento. Non esiste alcuna segnalazione molecolare nota tra sinapsi nel senso opposto che abbia questa proprietà. La LTP e la LTD modificano le sinapsi in modo locale e indipendente.

Crick (1989) suggerì, con una certa ironia, che se la backpropagation fosse il meccanismo biologico dell'apprendimento, dovremmo aspettarci di trovare molecole di segnalazione che percorrono gli assoni *all'indietro* trasportando informazione sulla modifica dei pesi anterogradi. Segnali retrogradi esistono — l'ossido nitrico (NO) e alcuni cannabinoidi endogeni sono messaggeri retrogradi che modulano il rilascio pre-sinaptico — ma essi operano su scala spaziale e temporale e con meccanismi di segnalazione incompatibili con il trasporto preciso di valori di pesi sinaptici.

Seconda obiezione: il problema del blocco modulante.

Nella backpropagation standard, la fase di aggiornamento dei pesi ($\mathbf{W}^l \leftarrow \mathbf{W}^l - \eta \cdot \delta^l (\mathbf{a}^{l-1})^T$) non può essere eseguita allo strato l finché il segnale di errore δ^l non è disponibile, il che richiede che tutto il passaggio backward dagli strati $L, L-1, \dots, l+1$ sia stato completato. In una rete con 100 strati e una latenza sinaptica di 1 ms per strato, questo imporrebbe un ritardo di 100 ms prima di qualunque aggiornamento allo strato 1 — durante il quale le sinapsi degli strati profondi devono rimanere invariate.

Il problema non è soltanto la latenza, ma la necessità di *congelare* lo stato della rete durante il passaggio backward. Se la rete continua a ricevere e processare input mentre il passaggio backward è in corso, le attivazioni $\mathbf{a}^l$ memorizzate durante il forward pass potrebbero non corrispondere più allo stato corrente della rete, rendendo i gradienti calcolati non più validi. In un sistema biologico che processa input continuamente, questa richiesta di congelamento è irrealizzabile.

Jaderberg et al. (2017) hanno proposto i *Synthetic Gradients* come soluzione ingegneristica a questo problema: uno strato aggiuntivo apprende a *predire* il segnale di gradiente che arriverà dallo strato superiore, permettendo all'aggiornamento dei pesi di procedere immediatamente senza attendere la propagazione completa del backward. Questo meccanismo, sebbene artificiale, suggerisce che la plausibilità biologica potrebbe essere recuperata se il cervello disponesse di circuiti capaci di *anticipare* i segnali di errore — una funzione che le connessioni predittive top-down del predictive coding potrebbero in linea di principio svolgere.

Terza obiezione: la non-località del segnale di errore.

L'aggiornamento sinaptico richiesto dalla backpropagation per la connessione tra il neurone j dello strato $\ell-1$ e il neurone i dello strato ℓ è:

$$\Delta W_{ji}^{\ell} = -\eta \cdot \delta_i^{\ell} \cdot a_j^{\ell-1}$$

Questo aggiornamento richiede due informazioni: l'attività pre-sinaptica $a_j^{\ell-1}$ (localmente disponibile alla sinapsi) e il segnale di errore δ_i^{ℓ} al neurone post-sinaptico. Il segnale δ_i^{ℓ} a sua volta è:

$$\delta_i^{\ell} = \sigma'(z_i^{\ell}) \cdot \sum_j W_{ji}^{\ell+1} \cdot \delta_j^{\ell+1}$$

Questo richiederebbe che il neurone i dello strato ℓ sappia: (a) quale sia il suo stato di attivazione corrente $\sigma'(z_i^{\ell})$ — questo è locale; (b) quali siano i pesi di tutte le sue connessioni uscenti $W_{ji}^{\ell+1}$ — plausibile se si ipotizza memoria locale; (c) quale sia il segnale di errore $\delta_j^{\ell+1}$ di tutti i neuroni j a cui è connesso allo strato superiore — non locale, richiede comunicazione da strati distanti.

Nel linguaggio della teoria del controllo, questo requisito equivale a chiedere che ogni componente del sistema abbia accesso allo *stato aggettivo* (*costate*) del sistema rispetto alla propria variabile interna — informazione che, nella soluzione classica delle equazioni di Pontryagin, si propaga all'indietro nell'intero sistema dagli stati finali verso quelli iniziali.

La non-località *spaziale* è dunque intrinseca alla struttura del problema dell'attribuzione del credito, non un artefatto dell'implementazione specifica della backpropagation. Qualunque soluzione biologicamente plausibile deve affrontarla, non aggirarla.

Quarta obiezione: la precisione aritmetica.

La backpropagation calcola gradienti con precisione in virgola mobile a 32 o 64 bit. Le sinapsi biologiche operano con una precisione intrinsecamente limitata: il numero discreto di vescicole pronte al rilascio, la variabilità poissoniana del rilascio del neurotrasmettitore, il rumore termico nelle correnti ioniche — tutti questi fattori introducono un livello di variabilità che equivale a lavorare con una risoluzione numerica di pochi bit.

Studi sulla capacità di memorizzazione delle sinapsi biologiche (Bhattacharya et al., 2017; Bhattacharya, 2019) suggeriscono che le sinapsi corticali abbiano una capacità effettiva di 4-5 bit per sinapsi in condizioni di massima plasticità, con rapide fluttuazioni spontanee che riducono la capacità informativa media a circa 1-2 bit. Questo contrasta con i tipici 32 bit dei pesi delle reti artificiali, una differenza di circa quattro ordini di grandezza in precisione numerica.

Tuttavia, questa obiezione deve essere contestualizzata. La discesa del gradiente stocastica (SGD) converge con successo anche usando gradienti computati con precisione a 8 bit (*quantized training*), e in alcuni casi anche a 4 bit. Il principio su cui si basa questa robustezza è che la direzione del gradiente è più importante della sua magnitudine esatta: se il segnale è nella direzione corretta (angolo $< 90^\circ$ con il vero gradiente), l'aggiornamento riduce la

perdita in aspettativa, indipendentemente dalla precisione numerica. La domanda biologicamente rilevante non è se le sinapsi possano implementare i calcoli con precisione arbitraria, ma se possano produrre aggiornamenti la cui *direzione media* sia sufficientemente allineata con il gradiente vero.

Quinta obiezione: la dicotomia addestramento/inferenza.

Questo punto è forse il meno fondamentale delle cinque obiezioni, ma il più visibile fenomenologicamente. Nelle reti artificiali, c'è uno stacco netto tra la fase di addestramento (pesi variabili, forward + backward pass, aggiornamenti del gradiente) e la fase di inferenza (pesi fissi, solo forward pass). Il cervello biologico non mostra tale dicotomia: le sinapsi sono plastiche anche durante il retrieval, anche durante compiti cognitivi attivi, anche durante il sonno — sebbene con intensità e meccanismi diversi.

Questa osservazione suggerisce che il cervello implementa qualcosa di più simile all'*online learning* — aggiornamento continuo dei parametri durante l'utilizzo — che all'apprendimento batch. I modelli di *continual learning* e di *meta-learning* in ambito artificiale tentano di approssimare questa proprietà, con risultati promettenti ma ancora lontani dalla flessibilità biologica.

La regola di Hebb e le sue formalizzazioni matematiche

La regola di Hebb (1949), nella sua formulazione originale, afferma che "quando il neurone A è abbastanza vicino ad attivare ripetutamente o persistentemente il neurone B, e partecipa al suo firing, avviene un processo di crescita o un cambiamento metabolico in una o entrambe le cellule tale che l'efficienza di A nel firing di B aumenta". Tradotta in termini matematici moderni:

$$\Delta W_{ij} = \eta \cdot a_i \cdot a$$

dove a_i e a sono rispettivamente le attivazioni del neurone post-sinaptico e pre-sinaptico. Questa regola è *puramente locale* e *non supervisionata*: dipende soltanto dall'attività delle due cellule connesse, senza alcun segnale esterno di errore.

Nella sua forma più semplice, la regola di Hebb è instabile: se a_i e a sono sempre positive (come nel caso di neuroni con firing rate non negativo), W_{ij} crescerà indefinitamente, portando a un'esplosione dei pesi. Diverse modificazioni sono state proposte per stabilizzarla:

Regola di Oja (1982). Aggiunge un termine di decadimento proporzionale al quadrato del peso:

$$\Delta W_{ij} = \eta (a_i a - a_i^2 W_{ij})$$

Oja dimostra che questa regola converge al vettore singolare principale dell'input (la prima componente principale), implementando un'analisi PCA locale. Il termine $-a_i^2 W_i$ corrisponde alla normalizzazione dell'output post-sinaptico, forzando $\|w_i\| \rightarrow 1$.

Regola BCM (Bienenstock, Cooper e Munro, 1982). Introduce una soglia di modifica θ_M che separa il regime LTD (bassa attivazione post-sinaptica) da quello LTP (alta attivazione):

$$\Delta W_i = \eta \cdot a_i \cdot \varphi(a_i, \theta_M)$$

dove $\varphi(a_i, \theta_M) < 0$ se $a_i < \theta_M$ e $\varphi(a_i, \theta_M) > 0$ se $a_i > \theta_M$. La caratteristica cruciale della BCM è che la soglia θ_M è *scorrevole*: essa si aggiorna in funzione dell'attività media storica del neurone,

$$d\theta_M/dt = (a_i^2 - \theta_M)/\tau_M$$

con τ_M costante di tempo. Questa soglia adattiva stabilizza la regola e, al contempo, implementa un meccanismo di omeostasi: se un neurone è mediamente iperattivo, la soglia si alza, favorendo la LTD; se è ipoattivo, la soglia si abbassa, favorendo la LTP. La BCM predice una dipendenza della plasticità sinaptica dal contesto storico di attivazione, confermata sperimentalmente nelle colonne corticali del gatto.

La STDP: derivazione e modelli computazionali

La *Spike-Timing-Dependent Plasticity* è la manifestazione sperimentale più diretta di un meccanismo di apprendimento locale con risoluzione temporale al millisecondo. La finestra di plasticità STDP standard è:

$$\begin{aligned} W(\Delta t) &= A_+ \exp(-\Delta t/\tau_+) \quad \text{se } \Delta t > 0 \quad (\text{pre} \rightarrow \text{post}, \text{LTP}) \\ W(\Delta t) &= -A_- \exp(\Delta t/\tau_-) \quad \text{se } \Delta t < 0 \quad (\text{post} \rightarrow \text{pre}, \text{LTD}) \end{aligned}$$

dove $\Delta t = t_{\text{post}} - t_{\text{pre}}$. I valori tipici nei neuroni piramidali della corteccia e dell'ippocampo sono: $A_+ \approx 0.01-0.05$, $A_- \approx 0.01-0.05$, $\tau_+ \approx 17$ ms, $\tau_- \approx 34$ ms. In molte sinapsi si osserva $A_-/A_+ \approx 1.05$, rendendo la regola leggermente asimmetrica verso la depressione — una proprietà necessaria per la stabilità del firing rate medio in reti con connettività ricorrente.

La STDP può essere implementata efficientemente mediante *tracce di elegibilità* (*eligibility traces*): ogni neurone mantiene una traccia della propria attività recente che decade esponenzialmente:

$$x_{\text{pre}}(t) = \sum_j : t_j^{\text{pre}} < t \exp(-(t - t_j^{\text{pre}})/\tau_+)$$

$$x_{\text{post}}(t) = \sum_i : t_i^{\text{post}} < t \exp(-(t - t_i^{\text{post}})/\tau_-)$$

L'aggiornamento sinaptico avviene in modo causale:

quando il neurone post-sinaptico genera uno spike: $\Delta W = +A_+ x_{\text{pre}}(t)$

quando il neurone pre-sinaptico genera uno spike: $\Delta W = -A_- x_{\text{post}}(t)$

Queste tracce possono essere aggiornate localmente a ogni spike, senza necessità di memorizzare l'intera storia temporale.

Connessione con la regola della catena temporale. Pfister e Gerstner (2006) hanno dimostrato che una regola STDP a tre fattori — che considera non solo coppie di spike ma anche triple (pre, post, post) e (pre, pre, post) — è in grado di apprendere sequenze temporali in modo equivalente all'algoritmo di backpropagation through time sotto opportune condizioni di codifica a rate. Questa non è equivalenza esatta, ma convergenza in aspettativa statisticamente analoga.

Instabilità della STDP in reti ricorrenti. La STDP standard in reti con connettività ricorrente tende a produrre dinamiche instabili: le sinapsi rinforzate aumentano il firing dei neuroni, che a loro volta rinforzano ulteriormente le sinapsi. Meccanismi di stabilizzazione sono necessari, tra cui: (a) omeostasi sinaptica (*synaptic scaling*, Turrigiano, 2008) — tutte le sinapsi di un neurone vengono scalate globalmente per mantenere un firing rate target; (b) potatura sinaptica competitiva — sinapsi deboli vengono eliminate se non rinforzate; (c) regole STDP asimmetriche con $A_- > A_+$ che producono una depressione netta in assenza di correlazioni. Questi meccanismi di stabilizzazione non sono descritti dalla backpropagation e costituiscono un livello aggiuntivo di organizzazione biologica senza controparte algoritmica evidente.

Predictive coding: derivazione variazionale completa

Il framework del predictive coding nella formulazione di Rao e Ballard (1999), poi esteso da Friston nel contesto del *free energy principle*, è l'architettura computazionale biologicamente plausibile più elaborata che abbia una connessione formale con la backpropagation. Presentiamo qui una derivazione completa.

Modello generativo. Si assume un modello generativo gerarchico gaussiano con L livelli. Il livello ℓ genera le attivazioni del livello $\ell-1$ attraverso una funzione $g^\ell: \mathbb{R}^{(n^\ell)} \rightarrow \mathbb{R}^{(n^{\ell-1})}$:

$$**r^{**\ell-1} \sim \mathcal{N}(g^\ell(**r^{**\ell}), \Sigma^\ell)$$

dove $\Sigma^\ell = \text{diag}(\sigma^{\ell 2})$ è la matrice di covarianza dell'errore di previsione al livello $\ell-1$, e r^ℓ è la rappresentazione al livello ℓ . Al livello più alto, un prior gaussiano:

$$**r^{**L} \sim \mathcal{N}(\mu_{\text{prior}}, \Sigma_{\text{prior}})$$

Inferenza variazionale. Dati le osservazioni $\mathbf{r}^0 = \mathbf{x}$, l'inferenza esatta calcola la posteriori $p(\mathbf{r}^1, \dots, \mathbf{r}^L \mid \mathbf{x})$, che è in generale intrattabile. L'inferenza variazionale approssima questa posteriori con una distribuzione fattorizzata $q(\mathbf{r}^1, \dots, \mathbf{r}^L) = \prod^l q(\mathbf{r}^l)$, scegliendo $q(\mathbf{r}^l) = \mathcal{N}(\mu^l, \Lambda^l)$. Minimizzando l'energia libera variazionale:

$$\begin{aligned} F &= \mathbb{E}_q [\log q(**r^{**}) - \log p(**r^{**0}, **r^{**1}, \dots, **r^{**L})] \\ &= \text{KL}[q(**r^{**}) \parallel p(**r^{**0} \mid **r^{**0})] - \log p(**r^{**0}) \end{aligned}$$

rispetto alle medie $\{\mu^l\}_{l=1}^L$ (le rappresentazioni inferite), si ottiene il sistema dinamico:

$$d\mu^l/dt = -\partial F/\partial \mu^l$$

Calcolando esplicitamente $\partial F/\partial \mu^l$ per il modello gaussiano:

$$\partial F/\partial \mu^l = (\Sigma^l)^{-1} **\epsilon^{**l} - ((\partial g^{l+1}/\partial \mu^l)^T (\Sigma^{l+1})^{-1} **\epsilon^{**l+1})$$

dove $\epsilon^l = \mu^l - g^{l+1}(\mu^{l+1})$ è l'errore di previsione al livello l (discrepanza tra rappresentazione e previsione top-down). Quindi:

$$d\mu^l/dt = -(\Sigma^l)^{-1} **\epsilon^{**l} + (\partial g^{l+1}/\partial \mu^l)^T (\Sigma^{l+1})^{-1} **\epsilon^{**l+1}$$

Questo sistema ha una struttura notevole: la variazione della rappresentazione al livello l dipende da: — l'errore di previsione locale ϵ^l (differenza tra rappresentazione corrente e previsione dal livello superiore); — l'errore di previsione al livello inferiore ϵ^{l+1} , pesato dal Jacobiano della funzione generativa.

Apprendimento dei parametri. Il gradiente della free energy rispetto ai parametri θ^l della funzione g^l è:

$$\partial F/\partial \theta^l = -(\partial g^l/\partial \theta^l)^T (\Sigma^l)^{-1} **\epsilon^{**l}$$

L'aggiornamento dei parametri è dunque:

$$d\theta^l/dt = (\partial g^l/\partial \theta^l)^T (\Sigma^l)^{-1} **\epsilon^{**l}$$

Questo è un aggiornamento locale: dipende soltanto dall'errore di previsione ϵ^l al livello corrente e dal Jacobiano di g^l rispetto ai parametri del livello corrente. Non richiede la propagazione esplicita di gradienti dall'uscita verso l'ingresso.

Equivalenza con la backpropagation. Millidge, Tschantz e Buckley (2020) dimostrano formalmente che, in architetture feedforward con funzioni generative lineari $g^l(\mathbf{r}^{l+1}) = \mathbf{W}^{l+1} \mathbf{r}^{l+1}$, e nel regime di convergenza completa delle equazioni dinamiche per μ^l (cioè $d\mu^l/dt = 0$ per ogni l prima di ogni aggiornamento dei parametri), gli aggiornamenti $\partial F/\partial \theta^l$ coincidono esattamente con i gradienti $\partial \mathcal{L}/\partial \mathbf{W}^l$ della backpropagation, dove $\mathcal{L} = (1/2) \|\mathbf{e}^0\|^2$ è l'errore quadratico di ricostruzione dello strato più basso.

La differenza è dunque computazionale, non matematica: la backpropagation calcola i gradienti in un singolo passaggio all'indietro, mentre il predictive coding li approssima mediante un processo iterativo di minimizzazione locale che richiede molte iterazioni per convergere.

Implementazione neurale. Rao e Ballard propongono che nel cervello: i neuroni piramidali degli strati profondi (V/VI) codifichino le rappresentazioni μ^l ; le cellule dell'errore di previsione (cellule superficiali degli strati II/III) codifichino $\mathbf{e}^l$; le connessioni feedforward trasmettano $\mathbf{e}^l$ verso i livelli superiori; le connessioni feedback trasmettano la previsione $g^l(\mu^{l+1})$ verso i livelli inferiori. Questa struttura anatomica funzionale è coerente con le osservazioni sull'asimmetria delle proiezioni corticali, sebbene non sia stata dimostrata a livello cellulare con sufficiente risoluzione.

Feedback alignment: analisi matematica della convergenza

Il risultato di Lillicrap et al. (2016) che le reti possono apprendere con matrici di feedback casuali fisse merita un'analisi della convergenza più precisa di quella offerta nella Parte II.

Sia $\mathbf{B}^l \in \mathbb{R}^{(n^l \times n^{l+1})}$ una matrice di feedback casuale fissa. Il segnale di errore approssimato è:

$$\delta^{l+1} = \mathbf{B}^{l+1} \delta^{l+2} \odot \sigma'(\mathbf{z}^{l+1})$$

anziché il segnale esatto $\delta^l = (\mathbf{W}^{l+1})^T \delta^{l+1} \odot \sigma'(\mathbf{z}^l)$. La direzione del vero gradiente rispetto ad $\mathbf{a}^l$ è $\mathbf{g}^l = (\mathbf{W}^{l+1})^T \delta^{l+1}$, mentre il segnale approssimato è $\tilde{\mathbf{g}}^l = \mathbf{B}^l \delta^{l+1}$. L'aggiornamento riduce la perdita se l'angolo θ tra $\mathbf{g}^l$ e $\tilde{\mathbf{g}}^l$ è inferiore a 90° :

$$\cos(\theta) = (\mathbf{g}^l \cdot \tilde{\mathbf{g}}^l) / (\|\mathbf{g}^l\| \cdot \|\tilde{\mathbf{g}}^l\|) > 0$$

Il meccanismo di feedback alignment è che durante l'addestramento, i pesi feedforward $\mathbf{W}^{l+1}$ si evolvono in modo da *allineare* la direzione del vero gradiente con quella del gradiente approssimato. Formalmente, si può mostrare che la componente di $\mathbf{W}^{l+1}$ lungo $\mathbf{B}^{lT}$ cresce, mentre la componente ortogonale a $\mathbf{B}^{lT}$ decade. Questo allineamento è guidato dalla stessa regola di aggiornamento:

$$\Delta \mathbf{W}^{l+1} = -\eta \delta^{l+1} (\mathbf{a}^l)^T$$

che tende a rafforzare le direzioni di $\mathbf{W}^{l+1}$ lungo cui $\mathbf{B}^T \boldsymbol{\delta}^{l+1}$ è allineato con il vero gradiente.

Nøkland (2016) estende il feedback alignment al *direct feedback alignment* (DFA), in cui ogni strato riceve il segnale di errore *direttamente* dall'uscita, non attraverso gli strati intermedi:

$$\mathbf{\delta}^{**l} = \mathbf{B}^{**l}_{\text{direct}} \mathbf{\delta}^{**L} \odot \sigma'(\mathbf{z}^{**l})$$

dove $\mathbf{B}^l_{\text{direct}} \in \mathbb{R}^{(n^l \times n^L)}$ è una matrice casuale fissa che mappa l'errore all'uscita direttamente allo strato l , saltando tutti gli strati intermedi. Questo elimina completamente la cascata di proiezioni backward e implementa una forma di apprendimento in cui ciascun livello apprende *indipendentemente* a minimizzare l'errore all'uscita — analoga alla struttura modulare dell'apprendimento nei sistemi biologici.

Target propagation e apprendimento modulare locale

La target propagation (TP) di LeCun (1986) e Bengio (2014) propone un approccio radicalmente diverso per aggirare il problema della non-località: invece di propagare gradienti, si propagano *obiettivi di attivazione*. L'idea è di definire, per ogni strato l , un target $\mathbf{t}^l \in \mathbb{R}^{(n^l)}$ — un'attivazione desiderata — e di usare questo target locale per calcolare un aggiornamento dei pesi che massimizzi la probabilità di raggiungerlo.

Target propagation esatta. Se la funzione di attivazione dello strato l avesse un'inversa esatta f^{l-1} , il target propagato all'indietro sarebbe:

$$\mathbf{t}^{**l} = f^{l-1}(\mathbf{t}^{**l+1}) \quad \text{per } l = L-1, L-2, \dots, 1$$

con $\mathbf{t}^L$ definito dalla funzione di perdita (per esempio, $\mathbf{t}^L = \mathbf{y}$ nella classificazione). L'aggiornamento dei pesi allo strato l minimizzerebbe la perdita locale:

$$\mathcal{L}^l = \|\mathbf{a}^{**l} - \mathbf{t}^{**l}\|^2$$

con gradiente rispetto a $\mathbf{W}^l$ calcolato localmente come nella backpropagation, ma usando il target locale $\mathbf{t}^l$ anziché il segnale di errore propagato dall'uscita.

Difference Target Propagation (DTP). Poiché le funzioni di attivazione delle reti neurali non sono generalmente invertibili, Lee et al. (2015) propongono di apprendere funzioni inverse approssimate $\mathbf{g}^l: \mathbb{R}^{(n^l)} \rightarrow \mathbb{R}^{(n^{l-1})}$, addestrate separatamente per ogni strato a soddisfare:

$$\mathbf{g}^{**l}(\mathbf{f}^{**l}(\mathbf{a}^{**})) \approx \mathbf{a}^{**}$$

Il target propagato è allora:

$$\mathbf{t}^{l+1} = \mathbf{a}^{l+1} + \mathbf{g}^{l+1}(\mathbf{t}^{l+1}) - \mathbf{g}^{l+1}(\mathbf{a}^{l+1})$$

dove il termine correttivo $\mathbf{g}^l(\mathbf{a}^{l+1}) - \mathbf{g}^l(\mathbf{t}^{l+1})$ compensa l'errore di approssimazione dell'inversa. Si può dimostrare che nel limite di inversione esatta, $\mathbf{t}^l \rightarrow \mathbf{a}^l + \mathbf{J}^{l-1}(\mathbf{t}^{l+1} - \mathbf{a}^{l+1})$, dove $\mathbf{J}^l$ è il Jacobiano di $\mathbf{f}^l$ — un aggiornamento che approssima il passo di Newton rispetto alla perdita locale. Millidge et al. (2020) hanno dimostrato una connessione formale tra DTP e predictive coding: entrambi possono essere visti come minimizzazioni di una funzione di energia distanziata dallo stato corrente, con la differenza che DTP usa distanze euclidee mentre il predictive coding usa distanze pesate dalla precisione.

Il modello bicompartimentale e burstprop: derivazione

Il modello di Sacramento et al. (2018), ribattezzato *burstprop*, è forse la proposta biologicamente più concreta tra quelle che approssimano la backpropagation. Si basa sulla struttura biofisica dei neuroni piramidali corticali, che presentano due zone di integrazione elettricamente distinte: il compartimento somatico/basale (riceve input dal livello immediatamente superiore) e il compartimento apicale (riceve input feedback dai livelli più alti).

Modello formale. Ogni neurone i allo strato ℓ ha due variabili di stato: l'attività somatica a_i^l e l'attività apicale u_i^l . Le equazioni del modello sono:

$$\begin{aligned} \tau \frac{da_i^l}{dt} &= -a_i^l + \sigma(\sum_j W_{ji}^l a_j^{l-1}) & (\text{input bottom-up}) \\ \tau \frac{du_i^l}{dt} &= -u_i^l + \sum_j Y_{ji}^l a_j^{l+1} & (\text{input top-down}) \end{aligned}$$

Il tasso di burst è modellato come:

$$\text{burst_rate}_i^l = \sigma(a_i^l + u_i^l) - \sigma(a_i^l)$$

che misura l'incremento di attività indotto dall'input apicale rispetto all'attività basale di default. La regola di plasticità è:

$$\Delta W_{ji}^l = \eta \cdot \text{burst_rate}_i^l \cdot a_j^{l-1}$$

Questo è un aggiornamento puramente locale: burst_rate_i^l è una quantità del neurone post-sinaptico i , e a_j^{l-1} è l'attività del neurone pre-sinaptico j .

Equivalenza con la backpropagation. Sacramento et al. dimostrano che se i pesi feedback $\mathbf{Y}^l$ soddisfano:

$$Y_i^{l+1} = \sigma^{l+1}(z_i^{l+1}) \cdot W^{l+1}_{ji}$$

(cioè la matrice di feedback è $\mathbf{Y}^l = \text{diag}(\sigma^{l+1}) \mathbf{W}^{l+1}$), allora:

$$\text{burst_rate}_i^l \approx \sigma^l(z_i^l) \cdot \sum_j Y_i^{l+1} a_j^{l+1} = \sigma^l(z_i^l) \cdot \sum_j W^{l+1}_{ji} \sigma^{l+1}(z_j^{l+1}) \cdot \dots$$

Il calcolo si propaga ricorsivamente producendo esattamente la formula δ_i^l della backpropagation. La condizione sui pesi feedback è tuttavia ancora non locale — richiede che Y dipenda da W — ma Sacramento et al. mostrano che una versione non supervisionata in cui $\mathbf{Y}^l$ viene appreso separatamente per invertire la funzione feedforward produce risultati comparabili.

Equilibrium propagation: il caso continuo e la connessione con i sistemi fisici

Scellier e Bengio (2017) descrivono l'equilibrium propagation (EP) come un framework per l'apprendimento in sistemi fisici con una funzione di energia. Generalizziamo la presentazione della Parte II con una trattazione più rigorosa.

Definizione. Sia $E(\mathbf{s}, \boldsymbol{\theta}; \mathbf{x})$ una funzione di energia differenziabile dove $\mathbf{s}$ sono le attività neuronali, $\boldsymbol{\theta}$ i parametri e $\mathbf{x}$ l'input. L'equilibrio libero $\mathbf{s}^0(\mathbf{x})$ minimizza l'energia totale:

$$\mathbf{s}^{**0} = \arg \min_{\mathbf{s}^{**}} E(\mathbf{s}^{**}, \boldsymbol{\theta}^{**}; \mathbf{x}^{**})$$

La fase perturbata (*nudged phase*) aggiunge un termine proporzionale a una funzione di costo $C(\mathbf{s}, \mathbf{y})$:

$$\mathbf{s}^{**\beta} = \arg \min_{\mathbf{s}^{**}} [E(\mathbf{s}^{**}, \boldsymbol{\theta}^{**}; \mathbf{x}^{**}) + \beta \cdot C(\mathbf{s}^{**}, \mathbf{y}^{**})]$$

Teorema centrale. Scellier e Bengio dimostrano che:

$$\partial \mathcal{L} / \partial \boldsymbol{\theta} = \lim_{\beta \rightarrow 0} \{ (1/\beta) [\partial E(\mathbf{s}^{**\beta}, \boldsymbol{\theta}) / \partial \boldsymbol{\theta} - \partial E(\mathbf{s}^{**0}, \boldsymbol{\theta}) / \partial \boldsymbol{\theta}]$$

dove $\mathcal{L} = C(\mathbf{s}^0, \mathbf{y})$ è la funzione di perdita valutata all'equilibrio libero. La prova usa il *teorema della funzione implicita* sull'equilibrio e la differenziabilità dell'energia:

Dimostrazione. Poiché $\mathbf{s}^0$ minimizza E , $\partial E(\mathbf{s}^0, \boldsymbol{\theta}) / \partial \mathbf{s} = 0$. Differenziando rispetto a β e valutando in $\beta = 0$:

$$d\mathbf{s}^{**\beta} / d\beta|_{\beta=0} = -(\partial^2 E / \partial \mathbf{s}^{**} \partial \mathbf{s}^{**T})^{-1} \cdot (\partial C / \partial \mathbf{s}^{**})$$

La variazione dell'energia rispetto ai parametri:

$$d/d\beta [\partial E(s^{\beta}, \theta)/\partial \theta]_{\{\beta=0\}} = (\partial^2 E / \partial s^{\beta} \partial \theta)^T \cdot (d s^{\beta} / d\beta)_{\{\beta=0\}}$$

D'altra parte, differenziando $\mathcal{L} = C(s^0, y)$ rispetto a θ e usando la condizione di primo ordine:

$$\partial \mathcal{L} / \partial \theta = (\partial C / \partial s^{*0})^T \cdot \partial s^{*0} / \partial \theta = -(\partial C / \partial s^{**})^T \cdot (\partial^2 E / \partial s^{**} \partial s^{**T})^{-1} \cdot (\partial^2 E / \partial s^{**} \partial \theta)$$

che coincide con il limite calcolato sopra, completando la dimostrazione. Il risultato è notevole perché non richiede mai la propagazione esplicita di un segnale di errore: il gradiente emerge dalla *differenza* tra due stati di equilibrio, che può essere misurata localmente a ogni sinapsi semplicemente confrontando l'attività delle due fasi.

Three-factor learning rules: formalizzazione e biologia molecolare

La regola delle tre fattori rappresenta la sintesi più convincente tra l'apprendimento per rinforzo e la plasticità sinaptica locale. La forma generale è:

$$\Delta W_{ij}(t) = \eta \cdot e_i(t) \cdot M(t)$$

dove $e_i(t)$ è la *traccia di eligibilità* sinaptica locale, accumulata sulla base dell'attività pre- e post-sinaptica, e $M(t)$ è il segnale di modulazione globale.

Traccia di eligibilità. In presenza di STDP, la traccia di eligibilità è:

$$e_i(t) = \int_{-\infty}^t K(t-s) \cdot a_{\text{pre}}(s) \cdot a_{\text{post}}(s) ds$$

dove $K(\tau)$ è la finestra temporale STDP. In pratica, la traccia decade esponenzialmente tra eventi di spike:

$$de_i/dt = -e_i/\tau_e + \text{STDP_update}$$

con τ_e tipicamente dell'ordine di secondi — una scala temporale che "congela" temporaneamente il ricordo dell'attività correlata recente in attesa di un segnale di rinforzo.

Segnale di modulazione. $M(t)$ è tipicamente il segnale dopaminergico RPE, ma può includere anche acetilcolina (segnale di novità/attenzione), noradrenalina (stress, incertezza) e serotonina (stati affettivi a lungo termine). La separazione temporale tra la formazione della traccia (attività neuronale, scala dei ms) e il segnale di modulazione (dopamina, scala dei secondi) risolve elegantemente il problema dell'attribuzione del credito temporale: la traccia

"ricorda" quali sinapsi erano attive durante un evento comportamentale, e il successivo segnale dopaminergico decide se rinforzarle o deprimerle.

Limitazione fondamentale. La three-factor rule risolve l'attribuzione del credito temporale ma non quello spaziale (in profondità). Il segnale dopaminergico è uguale per tutte le sinapsi in una regione cerebrale, indipendentemente dal loro contributo specifico all'errore. Non porta informazione sulla direzione del gradiente rispetto ai singoli pesi. Questa limitazione è intrinseca alla diffusività volumetrica della dopamina e non può essere superata semplicemente moltiplicando il segnale per tracce locali.

Parte III — L'evidenza sperimentale e i modelli computazionali vincolati biologicamente

Metodologie di indagine: registrazioni multielettrodo, calcio-imaging e connettomia

Il confronto tra teorie computazionali dell'apprendimento e substrato biologico richiede metodologie in grado di misurare l'attività di popolazioni neuronali con risoluzione spaziale e temporale sufficiente, e di quantificare le modifiche sinaptiche con precisione molecolare. Gli ultimi dieci anni hanno visto progressi metodologici straordinari su entrambi i fronti.

Le sonde *Neuropixels* (Jun et al., 2017) permettono la registrazione simultanea di centinaia di neuroni in una colonna corticale di 10 mm di profondità, con risoluzione temporale al sub-millisecondo. Le sonde di seconda generazione (*Neuropixels 2.0*) portano la densità di elettrodi a 5120 siti per sonda, consentendo la registrazione di migliaia di neuroni nella stessa sessione. Combinando più sonde in diverse aree corticali, è oggi possibile registrare simultaneamente popolazioni neuronali in strutture gerarchicamente distanti — per esempio V1, V4, IT e corteccia prefrontale nella scimmia — durante l'esecuzione di compiti cognitivi complessi.

Il calcio-imaging a due fotoni con indicatori fluorescenti geneticamente codificati (GCaMP6, jGCaMP8) permette la registrazione dell'attività di decine di migliaia di neuroni su aree corticali superficiali, con risoluzione cellulare. Sebbene la risoluzione temporale (decine di millisecondi) sia inferiore a quella dell'elettrofisiologia, il calcio-imaging consente campionamenti ripetuti della stessa popolazione cellulare nel corso di sessioni di apprendimento estese — giorni o settimane — permettendo di monitorare direttamente come le rappresentazioni neuronali si modificano durante l'acquisizione di una competenza. Ziv et al. (2013) hanno usato questa tecnica per mostrare che le rappresentazioni delle cellule di posto nell'ippocampo derivano da un sottoinsieme stabile ma continuamente sostituito di cellule attive, suggerendo una dinamica di apprendimento che non corrisponde alla semplice modificazione incrementale dei pesi esistenti, ma coinvolge la selezione e il reclutamento di nuove cellule.

Sul versante della connettomia, la *serial section electron microscopy* (ssEM) combinata con algoritmi di segmentazione automatica ha reso possibile la ricostruzione tridimensionale di circuiti neuronali a scala nanometrica. La collaborazione FlyWire ha completato nel 2023 il connettoma dell'intero cervello di *Drosophila melanogaster* — 140.000 neuroni e circa 50 milioni di sinapsi — mentre i progetti H01 (Human Cortex) e MICrONS hanno ricostruito porzioni di corteccia di topo a risoluzione sinaptica (200-300 μm^3 di volume). Questi dataset permettono di misurare la distribuzione delle efficacie sinaptiche, la struttura dei motivi di connettività locale e la simmetria o asimmetria delle connessioni reciproche — dati direttamente rilevanti per il problema della simmetria dei pesi.

I dati connettomici disponibili mostrano in modo univoco che la distribuzione delle efficacie sinaptiche nelle connessioni feedforward e feedback tra due aree corticali non mostra correlazione significativa. Egger et al. (2021) hanno esaminato le connessioni reciproche tra neuroni piramidali di area V1 e V2 del topo e non hanno trovato nessuna tendenza alla simmetria: la forza della connessione da A a B era statisticamente indipendente dalla forza della connessione da B ad A. Questo risultato, per quanto atteso sulla base delle obiezioni teoriche, costituisce una confutazione diretta e sperimentale della simmetria dei pesi come proprietà biologica.

Rappresentazioni gerarchiche: dal confronto CNN-corteccia ai modelli vincolati

Il programma di ricerca inaugurato da Yamins et al. (2014) e proseguito da DiCarlo, Yamins, Zador e collaboratori rappresenta l'approccio sistematicamente più ambizioso per confrontare rappresentazioni artificiali e biologiche.

Il metodo della Representational Similarity Analysis. Kriegeskorte, Mur e Bandettini (2008) propongono di confrontare le rappresentazioni di un insieme di stimoli in due sistemi (un'area cerebrale e un modello computazionale) attraverso le rispettive matrici di dissimilarità. Per un set di N stimoli, la matrice di dissimilarità *Representational Dissimilarity Matrix* (RDM) è una matrice $N \times N$ in cui l'elemento (i,j) misura la distanza tra le risposte agli stimoli i e j nello spazio delle rappresentazioni. Questa distanza è tipicamente la distanza di correlazione $1 - \rho$, dove ρ è il coefficiente di correlazione di Pearson tra i vettori di risposta ai due stimoli.

La corrispondenza tra due sistemi si misura come la correlazione di Spearman (o Kendall's τ) tra le loro RDM, dopo aver vettorializzato la parte triangolare superiore. Questa metrica ha la proprietà di essere invariante a trasformazioni lineari dei vettori di risposta, catturando la geometria relativa delle rappresentazioni piuttosto che le loro magnitudini assolute. L'approccio RSA ha il vantaggio cruciale di essere applicabile a sistemi con spazi di rappresentazione di dimensionalità diversa — per esempio neuroni con firing rate e unità con attivazioni float32 — permettendo il confronto diretto senza proiezioni arbitrarie.

Risultati del confronto CNN-corteccia visiva. Yamins e DiCarlo (2016) mostrano che, su un set di 1960 immagini di oggetti naturali fotografati in condizioni variate (illuminazione,

punto di vista, sfondo), la correlazione tra le RDM delle CNN e quelle delle aree IT della scimmia raggiunge $r \approx 0.55\text{--}0.65$ per i modelli migliori, contro $r \approx 0.3\text{--}0.4$ per i modelli precedenti (filtri di Gabor, SIFT, V1 modellato). Gli strati intermedi delle CNN predicono meglio V4, mentre gli strati profondi predicono meglio IT. Questa progressione gerarchica è robusta rispetto alla specifica architettura della CNN e all'insieme di stimoli usato.

Schrimpf et al. (2020) estendono questo confronto ai modelli di linguaggio, mostrando che gli strati intermedi dei transformer (BERT, GPT-2) predicono l'attività della corteccia temporale superiore e di Broca durante la lettura di frasi in misura comparabile a quella delle CNN per la visione — un risultato che suggerisce principi organizzativi comuni alle gerarchie sensoriali di diverso dominio.

Critiche metodologiche. Nonostante i risultati impressionanti, questa linea di ricerca ha ricevuto critiche metodologiche importanti. Storrs e Kriegeskorte (2019) sottolineano che le CNN addestrate su ImageNet hanno una struttura dei dati di addestramento (milioni di immagini fotografiche etichettate) radicalmente diversa dall'input sensoriale biologico (flusso visivo continuo, non etichettato, con dipendenze temporali). Il fatto che le loro rappresentazioni correlino con quelle della corteccia potrebbe riflettere una convergenza verso soluzioni ottimali per lo stesso tipo di compito (riconoscimento di oggetti naturali), indipendentemente dal meccanismo di apprendimento.

Klindt et al. (2019) mostrano che reti addestrate con predictive coding sui dati grezzi di un simulatore di ambiente naturale producono rappresentazioni delle aree V1 e V2 statisticamente più simili alle risposte neuronali di queste aree rispetto alle CNN addestrate con supervisionamento su ImageNet — specialmente per quanto riguarda la struttura spaziotemporale delle risposte e la soppressione centrosurround. Questo risultato è significativo perché suggerisce che il meccanismo di apprendimento, non solo il tipo di compito, influisce sulla forma delle rappresentazioni sviluppate.

Dinamiche neurali durante l'apprendimento: registrazioni in tempo reale

Se le registrazioni statiche mostrano *cosa* le reti biologiche rappresentano, le registrazioni dinamiche durante l'apprendimento mostrano *come* quelle rappresentazioni si modificano nel tempo — informazione più direttamente rilevante per il confronto tra algoritmi di apprendimento.

Apprendimento della struttura sequenziale. Pastalkova et al. (2008) e successivamente Dragoi e Tonegawa (2011) hanno registrato cellule di posto nell'ippocampo di ratto durante il tracciamento di percorsi su un labirinto lineare. Nella fase di pregiooco (*pre-play*), prima che il ratto abbia esplorato il labirinto, le cellule di posto già mostrano sequenze di attivazione nella fase di quiet wakefulness che corrispondono alle sequenze che verranno poi osservate durante l'esplorazione effettiva. Questo suggerisce che il cervello *anticipa* la struttura dell'ambiente mediante generazione offline di schemi di attività — un processo che

assomiglia alla generazione di dati sintetici in un modello generativo, più che all'aggiornamento reattivo dei pesi in risposta a errori osservati.

Formazione di cellule di posto durante esplorazione nuova. Frank et al. (2004) hanno mostrato che nel topo le cellule di posto in CA1 sviluppano campi di risposta stabili entro pochi minuti dall'ingresso in un ambiente nuovo, con modifiche sinaptiche rapide che avvengono entro la prima traversata completa del labirinto. La velocità di questo apprendimento — incompatibile con la lenta convergenza della backpropagation standard — è spiegata dalla BTSP (Bittner et al., 2017) descritta nella Parte II: plateau potentials indotti dall'input di CA3 sulle cellule di CA1 durante la traversata producono una LTP massiccia di centinaia di sinapsi in modo quasi istantaneo, codificando la sequenza spaziale con un'efficienza campionaria (di fatto, *one-shot*) irraggiungibile dalla discesa del gradiente.

Modifiche corticali durante apprendimento motorio. Shenoy, Sahani e Churchland (2013) hanno registrato popolazioni di neuroni nella corteccia motoria primaria (M1) e premotoria (PMd) di scimmia durante l'apprendimento di sequenze motorie nuove. Nei primi minuti di addestramento, la struttura delle attività di popolazione mostra una rapida riorganizzazione geometrica nello spazio delle rappresentazioni — i *neural manifolds* che descrivono lo spazio delle traiettorie possibili si riconfigurano verso strutture che minimizzano l'interferenza tra i movimenti nuovi e quelli già appresi. Questa riorganizzazione geometrica, analizzata mediante PCA sulle attività di popolazione, è incompatibile con semplici modifiche monotone dei pesi sinaptici (come ci si aspetterebbe dalla discesa del gradiente), e suggerisce un processo di riorganizzazione strutturale più profonda, possibilmente mediata da meccanismi di neuromodulazione che alterano simultaneamente la plasticità di interi circuiti.

Segnali di errore in corteccia: identificazione sperimentale

L'identificazione diretta di segnali di errore di previsione nella corteccia — distinti dai segnali di attività corrente e dai segnali di previsione — è una delle sfide metodologiche più ardue della neuroscienze computazionale, perché richiede di separare contributi sovrapposti nello stesso substrato neuronale.

MEG e EEG: segnatura temporale degli errori. Gli studi MEG/EEG sulla *mismatch negativity* (MMN) e sulla *P300* hanno documentato da decenni la risposta cerebrale a devianti inattesi in sequenze sensoriali regolari. La MMN, un componente negativo che emerge 150-250 ms dopo la presentazione di uno stimolo deviante, è considerata il correlato elettrofisiologico dell'errore di previsione nel sistema uditivo. Näätänen et al. (1978) ne identificarono la principale caratteristica: la MMN emerge automaticamente, senza necessità di attenzione diretta allo stimolo, indicando che il confronto tra input e previsione avviene in modo pre-attentivo, a livello di elaborazione sensoriale primaria. Questo è coerente con il predictive coding, in cui gli errori di previsione emergono autonomamente nell'architettura gerarchica senza richiedere un processo supervisivo.

fMRI e segnale BOLD. Murray et al. (2002) usano fMRI nel solco temporale superiore per mostrare che le regioni di ordine superiore presentano maggiore attività per stimoli

prevedibili rispetto a imprevedibili (inversione del gradiente rispetto a quanto atteso da semplice processamento bottom-up), coerentemente con la soppressione del feedback predittivo per stimoli già compresi. Sterzer et al. (2018) in una meta-analisi di 48 studi fMRI sul predictive coding mostrano che la dissociazione tra segnali di previsione (soppressione per stimoli attesi) e segnali di errore (attivazione per stimoli inattesi) è robusta e replicabile in corteccia visiva, uditiva e somatosensoriale, con un pattern gerarchico coerente con l'architettura di Rao-Ballard.

Registrazioni unitarie: cellule di errore e cellule di previsione. Rao e collaboratori (2019) hanno condotto un esperimento diretto in cui scimmie erano addestrate a predire la posizione di uno stimolo visivo sulla base di una sequenza temporale regolare. In V1, V2 e V4 hanno identificato due popolazioni distinte: cellule che rispondono più fortemente agli stimoli attesi (cellule di rappresentazione/previsione, negli strati profondi) e cellule che rispondono più fortemente agli stimoli inattesi (cellule di errore, negli strati superficiali). La ratio di queste due popolazioni cambia con la profondità gerarchica: nelle aree inferiori (V1, V2) predominano le cellule di errore, nelle aree superiori (V4, IT) predominano le cellule di rappresentazione — coerente con la previsione che le aree superiori generino previsioni e le inferiori le confrontino con l'input.

Tuttavia, è importante notare che questa dissociazione non è universale: alcuni studi (Keller e Mrsic-Flogel, 2018; Leinweber et al., 2017) mostrano che in V1 di topo la risposta alle perturbazioni predittive dipende fortemente dal tipo di compito e dalla storia di addestramento, con cellule che cambiano la loro classifica tra "errore" e "previsione" a seconda del contesto. Questo suggerisce che la distinzione non sia una proprietà anatomica fissa delle cellule, ma una proprietà funzionale dinamica che emerge dall'interazione con l'ambiente di addestramento.

Oscillazioni corticali e flussi di informazione gerarchica

La relazione tra oscillazioni corticali e flussi di informazione nelle gerarchie sensoriali è uno dei territori più fecondi e dibattuti della neuroscienze computazionale. Bastos et al. (2012, 2015) hanno investigato sistematicamente questa relazione mediante *Granger causality* spettrale tra registrazioni LFP multiarea in scimmia durante compiti visivi di attenzione.

La misura di Granger causality da un'area A verso un'area B a frequenza f misura in che misura la storia passata del segnale in A migliora la previsione del segnale corrente in B (al netto della storia passata di B stessa) alla frequenza f . Bastos et al. trovano un'asimmetria spettrale robusta:

— La Granger causality nella direzione feedforward (V1→V4, V4→TEO, TEO→TE) è dominata dalle frequenze gamma (40-80 Hz), con picchi tipicamente a 50-60 Hz; — La Granger causality nella direzione feedback (TE→TEO, TEO→V4, V4→V1) è dominata dalle frequenze alfa e beta (8-30 Hz), con picchi tipicamente a 10-15 Hz.

Questa dissociazione è interpretata nel framework del predictive coding come segue: i segnali feedforward ad alta frequenza trasportano errori di previsione (segnali che devono essere processati rapidamente e che contengono informazione nuova); i segnali feedback a bassa frequenza trasportano previsioni top-down (segnali che modulano lentamente l'attività degli strati inferiori sincronizzando le loro finestre di integrazione temporale).

Il meccanismo biofisico proposto (Mejias et al., 2016) si basa sulla differenza di scala temporale delle correnti di membrana nelle diverse popolazioni neuronali: i neuroni degli strati superficiali (II/III), che producono le proiezioni feedforward, hanno costanti di tempo brevi e generano risposta ad alta frequenza; i neuroni degli strati profondi (V/VI), che producono le proiezioni feedback, hanno costanti di tempo più lunghe (mediate in parte dalle correnti NMDA) e generano risposta a bassa frequenza.

La dissociazione spettrale ha trovato conferma in registrazioni nell'uomo mediante ECoG (electrocorticography) durante compiti di linguaggio e memoria (Leszczyński et al., 2020): le connessioni tra aree linguistiche e prefrontali mostrano la stessa asimmetria gamma/beta, con le connessioni feedforward (da corteccia temporale a prefrontale) dominate dal gamma e le feedback viceversa. Questo suggerisce che il principio sia preservato attraverso le specie e i domini sensoriali.

Il sistema dopaminergico: quantificazione dell'RPE e connessioni con il TD-learning

La relazione tra dopamina e TD-learning merita una trattazione quantitativa dettagliata, perché è il collegamento più diretto tra neurobiologia e un algoritmo computazionale specifico.

Il TD-learning e la convergenza. L'algoritmo $TD(\lambda)$ di Sutton (1988) aggiorna la funzione di valore $V(s)$ mediante:

$$V(s_t) \leftarrow V(s_t) + \alpha \delta_t e_t(s_t)$$

dove $\delta_t = r_{t+1} + \gamma V(s_{t+1}) - V(s_t)$ è il TD error e $e_t(s)$ è una traccia di elegibilità:

$$e_t(s) = \gamma \lambda e_{t-1}(s) + \mathbb{1}[s = s_t]$$

Con $\lambda = 0$ ($TD(0)$), l'aggiornamento usa solo la differenza temporale di un passo; con $\lambda = 1$, si recupera l'equivalenza con il Monte Carlo. La convergenza di $TD(\lambda)$ al valore vero V^* sotto politica fissa e tasso di apprendimento appropriato è garantita per $\lambda \in [0,1]$ in ambienti con stati discreti (Sutton, 1988).

Quantificazione sperimentale. Schultz, Dayan e Montague (1997) analizzano 160 neuroni dopaminergici nella substantia nigra pars compacta (SNc) e nell'area tegmentale ventrale

(VTA) di scimmia, durante un protocollo di condizionamento classico in cui un tono (CS) predice una goccia di succo (US). Le risposte documentate mostrano:

— Prima del condizionamento: burst fasici (aumento del firing rate da ~ 4 Hz a ~ 20 -30 Hz per 150-200 ms) all'occorrenza dell'US (ricompensa inattesa), con $\delta \approx r - 0 = r > 0$. — Dopo condizionamento completo: il burst si sposta temporalmente al CS (stimolo predittivo), con $\delta \approx r - V(\text{CS}) \approx 0$ all'US (perché $V(\text{CS}) \approx r$); pausa (*dip*) nel firing dell'US omesso, con $\delta \approx 0 - V(\text{CS}) < 0$. — Correlazione quantitativa: la magnitudine della risposta dopaminergica scala linearmente con il valore della ricompensa su un range di 0.05-0.5 ml di succo ($r^2 = 0.89$ nel paper originale).

Cohen et al. (2012) estendono queste osservazioni con registrazioni optogeneticamente identificate di neuroni dopaminergici, confermando che la risposta fasica codifica $\delta = r + \gamma V(s') - V(s)$ con precisione temporale al sub-secondo.

Striato e aggiornamento dei valori. Il segnale dopaminergico modulante è ricevuto principalmente dallo striato dorsale e ventrale, dove i neuroni spinosi medi (medium spiny neurons, MSN) esprimono recettori D1 (che aumentano l'eccitabilità e la LTP in risposta alla dopamina) o D2 (che diminuiscono l'eccitabilità e favoriscono la LTD). Questa differenziazione recettoriale implementa una forma di aggiornamento di valore differenziale: i neuroni D1 codificano il *critic* che aggiorna il valore verso l'alto in risposta a RPE positivi, mentre i neuroni D2 codificano la risposta a RPE negativi. Il modello "actor-critic" biologico (O'Doherty et al., 2004) mappa questa distinzione sul framework formale del TD-learning.

Limiti dell'analogia. La corrispondenza tra dopamina e TD error, per quanto straordinariamente precisa, non si estende immediatamente all'attribuzione del credito spaziale in reti profonde. Il TD error δ è un segnale scalare che informa sulla qualità globale della previsione del valore, non sulla contribuzione specifica di ogni sinapsi a quell'errore. Nelle applicazioni di deep reinforcement learning (DQN, A3C), l'attribuzione del credito spaziale è gestita dalla backpropagation sulla funzione Q o sulla policy network, con il TD error che fornisce soltanto il target di apprendimento. Nel cervello biologico, il meccanismo equivalente che disaggrega il segnale scalare dopaminergico nei contributi sinaptici specifici è ancora misterioso.

Consolidamento della memoria e replay durante il sonno: meccanismi e modelli

Il *memory consolidation* durante il sonno è uno dei processi di apprendimento biologico più documentati e, al contempo, tra i meno replicati nei sistemi artificiali. Wilson e McNaughton (1994) per primi documentano la *replay* sequenziale nell'ippocampo di ratto: sequenze di attivazione delle cellule di posto osservate durante l'esplorazione di un labirinto vengono riattivate durante il sonno NREM successivo, con una compressione temporale di fattore 6-7 (sequenze di decine di secondi vengono riprodotte in pochi secondi). Il replay avviene tipicamente durante i *sharp-wave ripples* (SWR) — brevi bursts di oscillazione ad alta frequenza

(100-250 Hz) sovrapposti ad un'onda di campo lenta — che coincidono con depolarizzazioni massicce delle cellule piramidali di CA1.

Diba e Buzsáki (2007) distinguono il *forward replay* (stesso ordine dell'esplorazione, comune nel sonno post-esperienza) dal *reverse replay* (ordine inverso, più frequente nella quiet wakefulness post-esperienza), e mostrano che il reverse replay potrebbe implementare qualcosa di funzionalmente simile a un passaggio backward: partendo dall'episodio più recente (spesso associato alla ricompensa) e propagando indietro nel tempo la modifica delle rappresentazioni degli stati precedenti — un processo analogo al TD(λ) con propagazione backward del valore lungo la traccia di eligibilità.

Il dialogo ippocampo-neocorteccia. Durante il sonno NREM, i SWR ippocampali sono sincronizzati con i *sleep spindles* talamocorticali (12-15 Hz) e con le *slow oscillations* (< 1 Hz) della neocorteccia. Questa sincronizzazione a tre vie — slow oscillations che coordinano spindles e SWR — è interpretata come il meccanismo attraverso cui l'ippocampo "scarica" i ricordi episodici recenti verso la neocorteccia per il consolidamento a lungo termine (Buzsáki, 1996; Diekelmann e Born, 2010).

Il modello computazionale di Kumaran, Hassabis e McClelland (2016) formalizza questo processo come un sistema di distillazione: l'ippocampo agisce da memoria rapida e ad alta capacità che codifica episodi singoli, e durante il replay trasferisce queste rappresentazioni a una rete neocorticale più lenta che le integra incrementalmente con le conoscenze pregresse, evitando l'interferenza catastrofica. Questo meccanismo biologico ha ispirato direttamente il *prioritized experience replay* di Schaul et al. (2016) nel deep reinforcement learning, in cui i ricordi più informativi (alta TD error) vengono campionati con maggiore frequenza durante l'addestramento offline — un parallelo con il fatto che gli SWR ippocampali sembrano preferenzialmente riattivare sequenze associate a ricompensa elevata.

Neuroni spinosi del cervelletto e apprendimento motorio di precisione

Il cervelletto è la struttura cerebrale in cui la relazione tra apprendimento sinaptico e correzione dell'errore è documentata con maggiore precisione biofisica. Il circuito cerebellare ha un'architettura sorprendentemente regolare: i neuroni di Purkinje (PC) dello strato corticale ricevono due tipi di input eccitatori — le *fibre parallele* (PF) dai granuli, che trasmettono informazione contestuale sugli stati sensorimotori, e le *fibre rampicanti* (CF) dell'oliva inferiore, che trasmettono segnali di errore di predizione del movimento.

La regola di plasticità proposta da Marr (1969) e Albus (1971) — e successivamente dimostrata sperimentalmente da Ito et al. (1982) — afferma che la co-attivazione di una fibra parallela e di una fibra rampicante produce una depressione a lungo termine (LTD) della sinapsi PF→PC. Formalmente:

$$\Delta w_{PF \rightarrow PC} = -\eta \cdot x_{PF} \cdot x_{CF} \cdot \sigma'(z)$$

dove x_{PF} è l'attività della fibra parallela pre-sinaptica, x_{CF} è l'attività della fibra rampicante (segnale di errore), e $\sigma'(z)$ è la sensibilità del potenziale di membrana della cellula di Purkinje. Questa è una three-factor rule nella sua forma più pura: le tre fattori sono l'attività pre-sinaptica (PF), l'attività post-sinaptica modulata dall'errore (CF), e la sensibilità locale (σ').

Il segnale trasmesso dalla CF è interpretato come un *teaching signal* che codifica l'errore tra il movimento pianificato e il movimento eseguito. Dean et al. (2010) hanno mostrato quantitativamente che la latenza e l'ampiezza della risposta delle CF sono proporzionali all'errore del movimento in maniera consistente con la previsione di un algoritmo di apprendimento per correzione dell'errore. La peculiarità del circuito cerebellare è che questo è un caso in cui l'architettura biologica è *più simile* alla backpropagation supervisionata di quanto non lo sia la corteccia cerebrale: le CF portano un segnale di errore specifico (il movimento sbagliato) verso le sinapsi responsabili (le PF che erano attive durante il movimento), producendo una modifica sinaptica proporzionale sia all'errore sia all'attività pre-sinaptica. Tuttavia, il segnale di CF è binario (si attiva o non si attiva) e limitato a un errore "di primo ordine" — non propaga un gradiente attraverso più strati di processamento.

Apprendimento nei trasformer biologici: la corteccia prefrontale e il working memory

La corteccia prefrontale (PFC) è coinvolta nelle forme più complesse e flessibili di cognizione: ragionamento, pianificazione, astrazione di regole, apprendimento few-shot. Il suo ruolo computazionale è spesso interpretato in termini di *working memory* — la capacità di mantenere attivamente rappresentazioni su scale temporali dell'ordine di secondi, oltre la finestra temporale della plasticità sinaptica rapida.

I modelli computazionali della PFC si ispirano sempre più alle architetture *transformer*: meccanismi di self-attention che permett l'integrazione flessibile di informazioni da posizioni diverse in una sequenza. Hahn et al. (2020) hanno mostrato che le risposte di popolazioni neuronali nella PFC laterale della scimmia durante compiti di memoria di lavoro sequenziale sono meglio predette da un transformer che da modelli ricorrenti classici (LSTM): il profilo temporale delle attivazioni, la struttura delle dipendenze tra step temporali, e la sensibilità all'ordine delle informazioni sono catturati meglio dall'architettura dell'attenzione che da quella ricorrente.

L'attenzione nei transformer non è addestrata con una regola locale; dipende dalla backpropagation per l'aggiornamento dei pesi delle matrici di query, key e value. Petrov et al. (2022) hanno proposto che nel cervello il meccanismo di attention sia implementato da neuromodulatori come l'acetilcolina, che modulano il guadagno sinaptico in modo selettivo in risposta alla salienza degli input. Questa proposta è biologicamente plausibile ma non ha ancora raggiunto la precisione formale necessaria per essere confrontata quantitativamente con i meccanismi di attenzione nei transformer.

Modelli di reti profonde biologicamente vincolate: un quadro sinottico

Concludiamo questa sezione tracciando un quadro dei principali modelli computazionali che tentano di soddisfare simultaneamente i vincoli biologici e i requisiti di prestazione su compiti di riferimento.

Il modello *CORnet* di Kumbhani et al. (2019) è una CNN ricorrente che aggiunge connessioni ricorrenti intra-area e proiezioni feedback inter-area all'architettura standard. CORnet ottiene una migliore correlazione con le risposte corticali di IT rispetto alle CNN feedforward pure, specialmente per stimoli difficili che richiedono più "cicli" di processamento gerarchico. L'addestramento rimane basato su backpropagation, ma l'architettura incorpora vincoli biologici sulla struttura delle connessioni.

Il modello *BioRNN* di Perez-Nieves e Bhatt (2021) implementa un predictive coding gerarchico con aggiornamenti sinaptici locali su immagini naturali, raggiungendo il 92% di accuratezza su CIFAR-10 con regole di apprendimento locali — non lontano dal 94-95% delle CNN addestrate con backpropagation, ma su architetture di profondità simile.

Il modello *Predictive Coding Network* di Whittington e Bogacz (2017, 2019) dimostra formalmente che un predictive coding gerarchico con una sola copia dei pesi (non simmetria dei pesi) apprende sequenze temporali equivalenti a quelle apprese da una LSTM addestrata con BPTT, purché la convergenza delle attività sia garantita prima di ogni aggiornamento dei pesi. Questo risultato è il più forte disponibile per la plausibilità biologica dell'apprendimento di sequenze, perché elimina sia il requisito della simmetria dei pesi sia quello del blocco modulante.

Insieme, questi modelli suggeriscono che la distanza tra "biologicamente plausibile" e "computazionalmente efficace" si è ridotta considerevolmente rispetto a una decade fa, sebbene rimanga ancora un divario non trascurabile nelle prestazioni su architetture molto profonde (>20 strati) e dataset di grande scala (ImageNet).

Parte IV — Sintesi critica, problemi aperti e prospettive future

Tre livelli di analisi e il rischio di confonderli

La struttura analitica proposta da David Marr nel 1982 in *Vision* — livello computazionale, livello algoritmico, livello implementazionale — è ancora oggi lo strumento concettuale più efficace per ordinare il dibattito sul rapporto tra backpropagation e neurobiologia. Vale la pena di applicarla sistematicamente ai temi trattati in questa monografia, prima di affrontare le questioni aperte.

Al *livello computazionale*, la domanda è: qual è la teoria di ciò che il sistema fa, e perché? Per la backpropagation in una rete neurale artificiale, la risposta è: il sistema minimizza una

funzione di perdita differenziabile rispetto ai propri parametri, mediante il calcolo esatto del gradiente. Per il cervello biologico, la risposta non è univoca e dipende dalla struttura del problema: nella corteccia visiva ventrale, il sistema potrebbe minimizzare un'energia di previsione gerarchica (free energy); nel cervelletto, minimizza un errore di previsione del movimento; nell'ippocampo, massimizza la fedeltà di encoding degli episodi. Non esiste un singolo obiettivo computazionale che descriva l'intero cervello, il che rende il confronto con la backpropagation (che ha un unico obiettivo ben definito) necessariamente parziale.

Al *livello algoritmico*, la domanda è: quale rappresentazione usa il sistema, e quale procedura segue? La backpropagation usa rappresentazioni reali (float32), vettori di segnale di errore propagati all'indietro, e la regola del gradiente disceso. I meccanismi biologici discussi — predictive coding, burstprop, feedback alignment, equilibrium propagation — usano rappresentazioni neurali (firing rates, burst rates, errori di previsione), dinamiche di settamento locale che convergono a soluzioni equivalenti al gradiente, e tracce di eligibilità per l'attribuzione del credito temporale. Le strutture algoritmiche sono diverse, ma convergono in determinati limiti allo stesso risultato.

Al *livello implementazionale*, le differenze sono radicali e non approssimabili: silicio vs. carbonio, virgola mobile vs. variabilità stocastica biochimica, sincronizzazione digitale vs. asincronismo distribuito, memoria globale vs. localismo sinaptico. A questo livello, la backpropagation e l'apprendimento biologico non potrebbero essere più distanti.

La confusione tra livelli — tipicamente il confondere una corrispondenza al livello computazionale con un'identità al livello algoritmico o implementazionale — è la fonte principale degli eccessi nella letteratura sia a favore sia contro l'ipotesi "il cervello usa la backpropagation". Entrambe le posizioni estreme, se formulate senza specificare a quale livello si riferiscono, sono insoddisfacenti.

Il problema dell'attribuzione del credito in profondità: stato dell'arte

La questione fondamentale che rimane irrisolta è se il cervello risolva il *credit assignment problem* nelle gerarchie corticali profonde — ossia se le sinapsi degli strati più profondi (V1, V2 nella gerarchia visiva; strati superiori del tronco in quella motoria) ricevano segnali correttivi proporzionali alla loro contribuzione all'errore comportamentale finale, con la specificità necessaria per guidare la loro modifica nella direzione corretta.

Argomenti a favore. Le evidenze per una forma di attribuzione del credito in profondità provengono principalmente da studi sullo sviluppo e sull'apprendimento percettivo. Schoups et al. (2001) mostrano che l'addestramento discriminativo su stimoli orientati produce modifiche sistematiche nelle proprietà di sintonizzazione di neuroni in V1 della scimmia — strati molto profondi nella gerarchia visiva ventrale — anche quando il compito discriminativo richiede chiaramente l'elaborazione di livelli superiori (V4, IT). Se le modifiche in V1 fossero semplicemente conseguenza di modifiche nell'input (l'animale guarda diversamente?), ci si aspetterebbe cambiamenti non specifici per orientamento; invece si osservano modifiche selettive per le orientazioni rilevanti per il compito, suggerendo un qualche tipo di segnale discendente specifico-di-compito che raggiunge V1.

Analogamente, studi di *top-down attention* (Motter, 1993; Treue e Martínez-Trujillo, 1999) mostrano che l'attenzione diretta a certe caratteristiche (colore, orientamento, velocità) modifica le risposte di neuroni di V1 in modo selettivo per quelle caratteristiche — una forma di modulazione discendente finemente accordata che potrebbe essere il substrato di segnali correttivi specifici.

Argomenti contro. La limitazione fondamentale degli studi sopra citati è che le modifiche osservate in V1 potrebbero essere mediate da connessioni feedback *non specificamente di errore*: basterebbe che i livelli superiori modificassero le previsioni top-down verso V1 — un processo previsto dal predictive coding indipendentemente dall'apprendimento — per produrre cambiamenti nelle proprietà di sintonizzazione senza richiedere un segnale di gradiente esplicito. Distinguere empiricamente tra "segnale feedback di errore proporzionale al gradiente" e "segnale feedback predittivo generico" richiede manipolazioni sperimentali molto precise che, ad oggi, non sono state realizzate.

Il *thought experiment* di Lillicrap e Simister (2019). Un modo per formalizzare questa questione è il seguente *thought experiment* computazionale: si prenda una rete di profondità L con architettura biologicamente ispirata (feedback alignment o predictive coding), si addestri sul compito visivo di riferimento, e si misuri la correlazione tra gli aggiornamenti sinaptici effettivamente realizzati allo strato 1 e il gradiente esatto calcolato dalla backpropagation allo stesso strato. Se questa correlazione è sistematicamente positiva ($\cos \theta > 0$, come dimostrato teoricamente per la feedback alignment in architetture semplici), allora il sistema *stava effettivamente risolvendo* il credit assignment problem, anche se non con la precisione della backpropagation. Simulazioni su modelli vincolati biologicamente mostrano correlazioni di 0.4-0.7 per le prime 100 epoche di addestramento, che poi crescono verso 0.8-0.9 per effetto del fenomeno di allineamento dei pesi descritto nella Parte II. Questo è il risultato più ottimistico disponibile, e suggerisce che la risoluzione approssimata del credit assignment problem potrebbe essere raggiunta naturalmente da meccanismi locali.

La sparsità come principio organizzativo dell'apprendimento

Un filo conduttore che emerge dalla biologia — e che la backpropagation standard non incorpora — è il principio della sparsità. A livello neuronale, la sparsità significa che solo una piccola frazione delle cellule è attiva per un dato input; a livello sinaptico, significa che solo un sottoinsieme delle connessioni possibili è presente e attivo.

Vantaggi computazionali della sparsità. Babadi e Sompolinsky (2014) dimostrano formalmente che, in reti associative di Hopfield, la capacità di memorizzazione scala con $(\log N)^{-1}$ per rappresentazioni dense (dove N è il numero di neuroni), ma scala linearmente con N per rappresentazioni con sparsità $f \rightarrow 0$ al crescere di N. In altre parole, la codifica sparsa permette di memorizzare esponenzialmente più pattern rispetto alla codifica densa, per lo stesso numero di sinapsi.

Sul fronte della discriminazione, Olshausen e Field (1997) e successivamente Arpit e Bengio (2019) mostrano che rappresentazioni sparse prodotte da una codifica che minimizza $\|x - D\alpha\|^2 + \lambda \|\alpha\|_1$ hanno una struttura geometrica (separazione tra cluster di punti nello spazio delle rappresentazioni) che facilita la classificazione con semplici regole lineari, anche per input non separabili linearmente nello spazio originale.

Apprendimento hebbiano con sparsità. La regola di Oja e le sue estensioni multi-neuroni (analisi delle componenti principali, ICA) producono rappresentazioni sparse quando le distribuzioni degli input sono supergaussiane (leptocurtiche), come avviene per le immagini naturali. Questo suggerisce che la sola esposizione a segnali naturali, senza supervisione esplicita, porta naturalmente verso rappresentazioni sparse — una spiegazione plausibile per la distribuzione delle cellule semplici di V1 senza richiedere algoritmi di apprendimento supervisionato.

Sparse autoencoders nei modelli di linguaggio. Un sviluppo recente di grande interesse è l'applicazione di sparse autoencoders per l'interpretabilità dei modelli linguistici (Anthropic, 2023; EleutherAI, 2024). Si addestra un autoencoder con perdita L1 per trovare una base sparsa delle attivazioni interne del transformer: si scopre che le "caratteristiche" estratte dalla base sparsa corrispondono a concetti semantici interpretabili (nomi propri, strutture grammaticali, sentimenti), mentre le attivazioni originali erano dense e non interpretabili direttamente. Questa linea di ricerca suggerisce che la sparsità potrebbe essere il formato naturale dell'informazione sia nei sistemi biologici sia nei transformer — ma che quest'ultimi, addestrati con backpropagation su token prediction, non ottimizzano esplicitamente per la sparsità e devono "estrarla" artificialmente.

Catastrophic forgetting e stabilità plastica: analisi approfondita

Il *catastrophic forgetting* rappresenta uno degli ostacoli più fondamentali nel confronto tra sistemi artificiali e biologici. La sua causa strutturale nella backpropagation è la seguente: i parametri $\theta \in \mathbb{R}^{(|\theta|)}$ dello spazio dei parametri sono condivisi tra tutti i compiti, e l'ottimizzazione per il compito B modifica i parametri in direzioni che possono distruggere le strutture nell'iperpiano dei parametri che codificavano il compito A.

Geometria del forgetting. Formalmente, dopo l'apprendimento del compito A con ottimale θ^* , la perdita sul compito A in un intorno di θ^* è approssimata dalla forma quadratica:

$$\Delta_A(\theta) \approx \Delta_A(\theta^*) + (1/2)(\theta - \theta^*)^T H_A (\theta - \theta^*)$$

dove H_A è la matrice hessiana della perdita A in θ^* . La matrice di informazione di Fisher F_A approssima H_A per modelli probabilistici:

$$F_A = \mathbb{E}[\nabla_{\theta} \log p(y^{**} | x^{**}, \theta) \nabla_{\theta} \log p(y^{**} | x^{**}, \theta)^T]$$

L'aggiornamento per il compito B sposta θ^* nella direzione $-\nabla_{\theta} \mathcal{L}_B$. Il forgetting di A è proporzionale alla componente di questo spostamento nella direzione degli autovettori di F_A con autovalori grandi — le direzioni dello spazio dei parametri più "importanti" per A.

L'algoritmo EWC di Kirkpatrick et al. (2017) minimizza:

$$\Delta_{\text{EWC}} = \Delta_B(\theta) + (\lambda/2) \sum_i F^A_i (\theta_i - \theta^*_{Ai})^2$$

dove F^A_i è la i -esima diagonale della matrice di Fisher per il compito A (approssimazione diagonale dell'hessiana). Questo penalizza le modifiche ai parametri più critici per A, equivalendo a proiettare la discesa del gradiente nello spazio ortogonale al sottospazio critico di A.

Analogia biologica. Il meccanismo biologico più direttamente analogo all'EWC è la *consolidazione sinaptica* mediata dalla PKM ζ (*Protein Kinase M zeta*), un'isoforma costitutivamente attiva di PKC che, una volta indotta dalla LTP, mantiene l'efficacia sinaptica rinforzata per ore o giorni resistendo alla depressione da LTD. Sacktor (2011) ha mostrato che l'inibizione della PKM ζ (mediante il peptide ZIP) produce una amnesia retrograda selettiva — cancellazione di memorie a lungo termine — senza effetti sulle sinapsi non consolidate. Questo suggerisce che le sinapsi consolidate siano effettivamente "protette" da modifiche successive, implementando biologicamente un analogo della penalizzazione EWC.

Replay e complementary learning. Un secondo meccanismo biologico contro il forgetting è il replay ippocampale descritto nella Parte III. Il modello di McClelland et al. (1995) propone che l'apprendimento proceda in due fasi: acquisizione rapida nell'ippocampo (con alto tasso di apprendimento η , a rischio di forgetting ma con alta fedeltà episodica) e consolidamento lento nella neocorteccia (con basso tasso di apprendimento, con basso rischio di forgetting ma minore risoluzione episodica). Il replay interleaved dei ricordi ippocampali durante il consolidamento neocorticale produce una forma di *mixed training* che previene il forgetting catastrofico: la neocorteccia "vede" sia i nuovi episodi sia i vecchi, e li integra in modo graduale.

L'implementazione artificiale più fedele di questo meccanismo è il *progressive neural network* di Rusu et al. (2016), che "congela" i parametri del compito precedente e aggiunge nuove colonne di neuroni per i compiti successivi, collegando le colonne attraverso *lateral connections*. Questo evita categoricamente il forgetting (le vecchie colonne sono immutabili) a costo di una crescita lineare del numero di parametri con il numero di compiti — un costo che la neocorteccia evidentemente non sostiene, grazie a meccanismi di compressione e generalizzazione che rimangono poco compresi.

Il dibattito sulla coscienza e il suo ruolo nell'apprendimento

La teoria dello *spazio di lavoro globale* (Global Workspace Theory, GWT) di Baars (1988), nella sua formulazione neurobiologica di Dehaene, Changeux e collaboratori (2001, 2011),

propone che la coscienza sia il meccanismo attraverso cui un'informazione locale — elaborata in un modulo specializzato — viene "trasmessa" a tutta la rete corticale, diventando disponibile per la competizione tra sistemi con obiettivi diversi. Il substrato anatomico del workspace globale è identificato con le connessioni di lunga portata della corteccia frontoparietale, i neuroni piramidali di strato V con assoni molto ramificati (*cortical neurons with long-range axons*), e le oscillazioni gamma sincronizzate su scala corticale larga.

GWT e attribuzione del credito. La proprietà più interessante della GWT nel contesto di questa monografia è la *diffusione di informazione*: quando un'informazione entra nel workspace globale, essa diventa disponibile a tutti i moduli che lo monitorano. Se questo workspace trasmette informazioni sull'*esito* (positivo o negativo) di un'azione, allora tutti i moduli che hanno contribuito a quell'esito possono ricevere questo segnale globale — non un gradiente specifico per ogni modulo, ma almeno un segnale di rinforzo. Combinato con tracce di elegibilità locali, questo schema è equivalente alla three-factor rule e potrebbe, in linea di principio, supportare forme di attribuzione del credito parzialmente specifiche se il segnale del workspace è sufficientemente informativo sul contributo di ciascun modulo.

IIT e complessità integrata. La *Integrated Information Theory* di Tononi (2004, 2014) propone che la coscienza sia identica all'informazione integrata φ di un sistema, definita come la quantità di informazione generata dal sistema "come un tutto" in eccesso rispetto alla somma dell'informazione generata dalle sue parti. φ è calcolabile (in linea di principio) come:

$$\varphi = \min_{\{ \text{partition } P \}} D(p(\mathbf{x} \mid \mathbf{x}_{-1}) \parallel \prod_i p(\mathbf{x}_i \mid \mathbf{x}_{-1}^{(i)}))$$

dove D è la divergenza KL e P è la bipartizione che minimizza l'informazione scambiata tra le due parti. Sistemi con alta φ hanno una struttura di dipendenze causali che non può essere decomposta in parti indipendenti — una proprietà che potrebbe essere rilevante per l'apprendimento: sistemi con alta φ potrebbero essere più capaci di attribuire credito a lungo raggio, perché le loro componenti non possono essere trattate come agenti indipendenti.

La connessione tra IIT e apprendimento gerarchico è ancora speculativa — le misure di φ scalano esponenzialmente con la dimensione del sistema, rendendo il suo calcolo esatto impraticabile per reti neurali biologiche — ma la direzione di ricerca è intellettualmente stimolante.

Meta-apprendimento: l'evoluzione come outer-loop dell'ottimizzazione

Una prospettiva che trasforma radicalmente il problema della plausibilità biologica è quella del *meta-apprendimento evolutivo*. Se consideriamo l'evoluzione come un processo di ottimizzazione outer-loop sulle strutture della rete (connettoma, distribuzione dei neurotrasmettitori, regole di plasticità), allora il risultato di questo processo — il cervello come lo osserviamo negli animali adulti — è una struttura già in parte ottimizzata per il tipo di ambiente e compiti che gli animali hanno affrontato nella loro storia evolutiva.

L'argomento di Zador. Zador (2019) sostiene che la gran parte del comportamento adattativo degli animali in condizioni naturali non emerge dall'apprendimento esperienziale durante la vita dell'individuo, ma è *innata*: codificata nella struttura del connettoma attraverso la selezione evolutiva. Il sistema visivo del topo sviluppa risposte orientamento-selettive in V1 anche in animali allevati al buio dalla nascita (Blakemore e Cooper, 1970; successivamente confermato da Hubel e Wiesel); il sistema uditivo degli uccelli canori codifica geneticamente la struttura generale del canto; i circuiti di controllo motorio dello spazio spinale sono funzionali nei neonati senza alcun apprendimento motorio. In tutti questi casi, la struttura del connettoma — non l'esperienza individuale — è il veicolo dell'adattamento.

In questa prospettiva, il problema computazionale che si pone non è "come ogni individuo apprende dalla scratch con meccanismi locali", ma "come l'evoluzione ha strutturato reti che già contengono soluzioni parzialmente valide, che l'esperienza individuale affina". Il ruolo della backpropagation (o dei suoi equivalenti biologici) sarebbe allora confinato a questo *affinamento* su un substrato pre-strutturato, non alla costruzione da zero di capacità cognitive complesse.

MAML e inizializzazioni evolutivamente ottimizzate. Il framework MAML di Finn et al. (2017) implementa computazionalmente questa idea: si cerca un'inizializzazione θ_0 dei pesi tale che poche iterazioni di gradient descent su un nuovo compito producano alta performance. L'inizializzazione ottimale θ_0 codifica, implicitamente, le strutture invarianti tra i compiti — analogo alle strutture evolutivamente conservate del connettoma. Nell'animale, θ_0 è determinato dall'evoluzione (su scala di milioni di anni), mentre negli esperimenti MAML è determinato da un meta-training su una distribuzione di compiti rappresentativa (su scale di ore/giorni di computazione).

Extrinsic vs. intrinsic learning signals. Schmidhuber (1991) ha proposto la distinzione tra segnali di apprendimento *estrinseci* (provenienti dall'ambiente, come ricompense o errori di predizione) e *intrinseci* (generati internamente dal sistema, come la riduzione della complessità di descrizione, la compressione, la curiosità). I segnali intrinseci sono computazionalmente analoghi ai termini di regolarizzazione nelle reti artificiali, ma nella biologia corrispondono a meccanismi come il *curiosity drive* e l'*intrinsic motivation*, mediati da neurotrasmettitori come la noradrenalina e l'acetilcolina in risposta alla novità. Sistemi che combinano segnali estrinseci (ricompensa) e intrinseci (novità, compressione) mostrano una generalizzazione molto superiore a sistemi che usano solo uno dei due — una lezione che i recenti sistemi di reinforcement learning (come RLHF applicato ai modelli linguistici) cominciano a incorporare.

La frontiera del neuromorphic computing

Il divario tra neurobiologia e intelligenza artificiale ha una dimensione non solo teorica ma anche ingegneristica: i circuiti integrati convenzionali (CPU, GPU) sono architetture di von Neumann ottimizzate per computazione sincrona e accurata, radicalmente diverse dal substrato biologico asincrono e stocastico. Il *neuromorphic computing* è il campo che tenta di

costruire hardware digitale o analogico ispirato all'architettura neuronale, con l'obiettivo di ottenere efficienza energetica comparabile al cervello biologico.

I chip neuromorfici come Intel Loihi (Davies et al., 2018), IBM TrueNorth (Merolla et al., 2014) e la piattaforma SpiNNaker di Manchester implementano modelli a spike (neuroni integrate-and-fire) con comunicazione asincrona, event-driven, senza clock globale. Intel Loihi supporta plasticità sinaptica on-chip con regole programmate (incluse varianti STDP e three-factor rules), permettendo l'apprendimento in-situ senza trasferimento dei dati su host esterno.

Sul piano dell'efficienza energetica, Loihi consuma circa 100 mW per milione di neuroni sintetici, mentre un GPU di pari capacità computazionale sincrona consuma diversi ordini di grandezza in più. Questa differenza riflette l'efficienza dell'architettura event-driven: i neuroni computano soltanto quando ricevono spike, mentre le GPU eseguono operazioni matriciali dense indipendentemente dalla sparsità degli input.

La limitazione attuale del neuromorphic computing è l'implementazione della backpropagation o dei suoi equivalenti biologici su questi substrati. L'equilibrium propagation è stata implementata su circuiti analogici (Scellier e Bengio, 2023), mostrando che la differenza di fase tra equilibrio libero e perturbato può essere calcolata direttamente sul chip senza propagare gradienti digitali. Il feedback alignment è naturalmente adatto all'implementazione su hardware neuromorfico perché la matrice di feedback casuale può essere codificata nel silicio al momento della fabbricazione, senza richiedere aggiornamenti. Questi risultati aprono la prospettiva di sistemi artificiali che apprendono con meccanismi biologicamente plausibili su hardware biologicamente ispirato — una convergenza che avrebbe implicazioni pratiche significative per l'efficienza energetica dell'intelligenza artificiale.

Le domande irrisolte: una mappa del territorio inesplorato

Concludere questa monografia senza indicare esplicitamente le domande ancora aperte sarebbe una reticenza intellettuale. Tre questioni emergono come particolarmente urgenti e non ancora aggredibili con i metodi disponibili.

Prima questione: la specificità dei segnali di modulazione. La three-factor rule richiede un segnale di modulazione globale $M(t)$ che moduli selettivamente le sinapsi eligibili. Ma nelle aree corticali associative, dove il contesto è complesso e i compiti sono multipli, la selettività anatomica della diffusione di dopamina (e degli altri neuromodulatori) potrebbe non essere sufficiente a indirizzare il segnale di modifica alle sole sinapsi rilevanti. Esiste un livello di specificità nel segnale dopaminergico che vada oltre la semplice diffusione volumetrica? Pubblicazioni recenti (Roitman et al., 2023) suggeriscono che la dopamina possa essere rilasciata in modo quasi sinaptico (*hotspot release*) in alcune regioni, con specificità sub-micronometrica. Se confermato su scala più ampia, questo cambierebbe radicalmente la stima della specificità del segnale di modifica e la sua capacità di implementare l'attribuzione del credito spaziale.

Seconda questione: la scala temporale dell'apprendimento gerarchico. Gli animali imparano competenze percettive complesse — riconoscimento di migliaia di categorie visive, comprensione del linguaggio, ragionamento analogico — nel corso di anni di sviluppo, con un'efficienza campionaria straordinariamente superiore a quella delle reti addestrate con backpropagation su dataset comparabili. È ipotizzabile che questa efficienza derivi non da un meccanismo di apprendimento più efficiente, ma da una struttura innata pre-ottimizzata dall'evoluzione (argomento di Zador), combinata con strategie di curriculum e di meta-apprendimento che le reti artificiali non implementano. Verificare questa ipotesi richiederebbe esperimenti di privazione sensoriale o di apprendimento in ambienti statisticamente distanti dall'ambiente evolutivo — esperimenti eticamente impraticabili negli umani ma possibili in modelli animali con architetture neurali semplificate.

Terza questione: l'emergenza delle strutture rappresentazionali. I cervelli biologici sviluppano rappresentazioni corticali altamente strutturate — mappe topografiche, colonne di dominanza oculare, moduli di riconoscimento dei volti — mediante processi di sviluppo che combinano istruzioni genetiche, attività spontanea prenatale, e apprendimento guidato dall'esperienza. I modelli computazionali attuali raramente incorporano questa dimensione sviluppativa: si parte da pesi inizializzati casualmente e si apprende in modo stazionario. L'idea che la struttura iniziale dei pesi (corrispondente al connettoma neonatale) contenga già informazione rilevante — e che l'apprendimento sia principalmente un affinamento di questa struttura — è probabilmente più vicina alla realtà biologica, ma richiede modelli computazionali che trattino lo sviluppo e l'apprendimento come processi distinti e interagenti.

Il gradiente come geometria dell'adattamento

Giunti al termine di questa trattazione, possiamo articolare la conclusione in modo più preciso di quanto fosse possibile nella versione iniziale. Il gradiente non è, in primo luogo, un algoritmo: è una *struttura geometrica*. Per qualunque funzione differenziabile $\mathcal{L}$ definita sullo spazio dei parametri $\mathbb{R}^{(|\theta|)}$, il gradiente $\nabla_{\theta} \mathcal{L}$ è il vettore campo che, in ogni punto, indica la direzione di massima crescita di $\mathcal{L}$. La minimizzazione di $\mathcal{L}$ mediante il seguire la direzione opposta — discesa del gradiente — è l'unica procedura di modifica locale dei parametri che garantisce la riduzione di $\mathcal{L}$ in aspettativa, per step sufficientemente piccoli.

Qualunque sistema che apprende — biologico o artificiale — e che possa essere descritto come la minimizzazione di un funzionale (anche approssimato, anche stocastico, anche implicito) dei propri parametri, *deve* in qualche senso approssimare la discesa del gradiente di quel funzionale. Non è una scelta: è un vincolo geometrico intrinseco al problema. La diversità tra gli algoritmi — backpropagation, predictive coding, equilibrium propagation, feedback alignment, STDP con modulazione dopaminergica — non è una diversità di principio, ma una diversità nel *grado di approssimazione* con cui ciascuno stima il gradiente e nella *struttura fisica* attraverso cui realizza questa stima.

In questa prospettiva, la domanda "il cervello usa la backpropagation?" perde il suo carattere di domanda binaria e assume la forma più ricca: in quale misura, attraverso quali meccanismi distribuiti, e con quale grado di approssimazione, il cervello biologico approssima il gradiente del funzionale che la selezione evolutiva e lo sviluppo ontogenetico hanno implicitamente definito per ciascun circuito neurale? Rispondere a questa domanda non richiede soltanto neuroscienze computazionale e ingegneria del machine learning, ma anche teoria evolutiva, biofisica molecolare, e — forse — una teoria della coscienza come meccanismo di coordinazione globale dell'apprendimento distribuito. È il tipo di domanda che attraversa i confini disciplinari, e questa caratteristica è, in ultima analisi, la misura più affidabile della sua profondità.

Parte V — Circuiti inibitori, neuromodulatori e calcolo dendritico

Il ruolo degli interneuroni: oltre la semplice inibizione

La letteratura sul confronto tra backpropagation e apprendimento biologico si concentra quasi invariabilmente sui neuroni eccitatori — le cellule piramidali che costituiscono circa l'80% dei neuroni corticali — trattando l'inibizione come un semplice meccanismo di controllo del guadagno. Questa semplificazione nasconde una delle architetture computazionali più sofisticate dell'intero sistema nervoso. Gli interneuroni GABAergici corticali non sono elementi passivi di soppressione: sono circuiti specializzati con proprietà di connettività, cinetica e plasticità radicalmente diverse dai neuroni piramidali, e svolgono un ruolo attivo nell'organizzazione dell'apprendimento gerarchico che non ha ancora trovato una corrispondenza sistematica nella letteratura sul deep learning.

La classificazione degli interneuroni corticali è tuttora dibattuta, ma una tassonomia funzionale ampiamente accettata distingue almeno tre tipi principali sulla base del bersaglio anatomico delle loro proiezioni sinapiche. Gli interneuroni *parvalbumin-positivi* (PV+), che includono le cellule basket e le cellule chandelier, innervano preferenzialmente il soma e i segmenti iniziali dell'assone delle cellule piramidali. Questa posizione anatomica li rende regolatori diretti del tasso di firing e della sincronia delle popolazioni piramidale: le cellule PV+ possono inibire precisamente e con bassissima latenza (< 1 ms) la generazione di potenziali d'azione, implementando una forma di controllo del guadagno quasi istantaneo. Le cellule *somatostatin-positive* (SST+), tra cui le cellule di Martinotti, innervano preferenzialmente i dendriti apicali delle cellule piramidali — esattamente il compartimento che, nel modello burstprop, riceve i segnali di feedback predittivi. Le cellule *VIP-positive* (VIP+) innervano preferenzialmente le altre categorie di interneuroni, generando un circuito di *dis-inibizione* in cui la loro attivazione rimuove l'inibizione esercitata da SST+ e PV+ sulle cellule piramidali.

Implicazioni per l'apprendimento gerarchico. Keller e Mrosovsky (2018), e successivamente Poort et al. (2022), hanno mostrato che durante l'apprendimento percettivo le cellule VIP+ in V1 di topo vengono attivate preferenzialmente dal feedback dalla corteccia motoria — segnale che indica che il topo *sta per muovere*, ovvero che l'animale ha aspettative

attive sul proprio campo visivo. L'attivazione delle VIP+ sopprime le SST+, che a loro volta disinibisce i dendriti apicali delle piramidal: in questo modo, il feedback motorio *apre una finestra di plasticità nei dendriti apicali proprio nei momenti in cui le previsioni dell'animale sono più precise* — implementando una forma di gate della plasticità che dipende dal contesto comportamentale. Il circuito VIP → SST → piramide realizza così una *modulazione contestuale della porta di apprendimento* che non ha ancora un omologo preciso nella backpropagation, ma che corrisponde funzionalmente all'idea di un segnale di modulazione $M(t)$ nella three-factor rule che è selettivamente attivo nei momenti di maggiore attendibilità predittiva.

Inibizione laterale e competition hebbiana. Un secondo ruolo fondamentale degli interneuroni è l'inibizione laterale tra popolazioni della stessa area corticale. Il meccanismo di *winner-take-all* (WTA), implementato da cellule PV+ che inibiscono le piramidal vicine in modo dipendente dall'attività, produce una competizione tra neuroni dello stesso livello gerarchico per la rappresentazione degli input. Questa competizione è il meccanismo alla base della formazione di rappresentazioni sparse: soltanto il neurone (o il piccolo sottoinsieme di neuroni) con la maggiore correlazione con l'input corrente viene autorizzato a firing, mentre gli altri vengono soppressi. Il risultato è una codifica k-sparse in cui esattamente k neuroni sono attivi per ogni input — analogamente al risultato dell'ottimizzazione della funzione:

$$\min_{\mathbf{a}} \|\mathbf{x} - \mathbf{D}\mathbf{a}\|^2 + \lambda \|\mathbf{a}\|_1$$

nella codifica sparsa di Olshausen e Field. L'inibizione laterale mediata da PV+ implementa biologicamente il termine di regolarizzazione ℓ_1 che produce sparsità nelle reti artificiali, ma mediante un meccanismo dinamico che non richiede la formulazione esplicita di un'energia da minimizzare — la competizione emerge spontaneamente dall'interazione locale tra cellule eccitatorie e inibitorie.

Neuromodulazione: la farmacologia del tasso di apprendimento

La three-factor rule discussa nelle parti precedenti identifica il segnale dopaminergico come il terzo fattore che modula la plasticità sinaptica oltre all'attività pre- e post-sinaptica. Ma il sistema neuromodulatorio del cervello è considerevolmente più ricco: almeno quattro sistemi distinti — dopaminergico, colinergico, noradrenergico e serotoninergico — proiettano diffusamente su tutta la corteccia con azioni complementari che, insieme, implementano un sistema di controllo adattivo del meta-apprendimento senza precedenti nei sistemi artificiali.

Il sistema colinergico e la selettività dell'apprendimento. Il nucleo basale di Meynert e il setto mediale proiettano fibre colinergiche su tutta la corteccia e sull'ippocampo. L'acetilcolina (ACh) è rilasciata in risposta a stimoli nuovi, imprevedibili o comportamentalmente rilevanti — una firma di *sorpresa* o *saliency* che è distinta dalla firma di *errore di previsione della ricompensa* codificata dalla dopamina. L'ACh agisce su recettori muscarinici M1 e nicotinici $\alpha 7$ sulle cellule piramidal, producendo una depolarizzazione

sostenuta e un aumento della conduttanza ai canali che amplifica la risposta alle correnti sinaptiche afferenti. L'effetto netto è un aumento del *signal-to-noise ratio* nelle risposte corticali: gli input forti diventano più forti e gli input deboli vengono soppressi.

Carriero et al. (2015) e Bear e Singer (1986) avevano mostrato che la deplezione colinergica impedisce la plasticità corticale durante i periodi critici dello sviluppo visivo — la formazione delle colonne di dominanza oculare in V1 richiede integro il sistema colinergico. Questo risultato indica che l'ACh non è soltanto un amplificatore del segnale, ma un *gate della plasticità*: in sua assenza, anche la co-attivazione sinaptica intensa non produce modifiche durature. In termini della three-factor rule, l'ACh modula la finestra di elegibilità — il periodo durante il quale la traccia sinaptica è "aperta" a ricevere il segnale di modifica — piuttosto che la direzione della modifica stessa.

Il ruolo dell'ACh come modulatore dell'attenzione è stato formalizzato da Yu e Dayan (2005) nel framework bayesiano: l'ACh codifica l'*expected uncertainty* — l'incertezza prevista sull'affidabilità delle previsioni attuali, derivante dalla variabilità intrinseca degli stimoli noti — mentre la noradrenalina codifica l'*unexpected uncertainty* — l'incertezza derivante da cambiamenti improvvisi nella struttura dell'ambiente. In termini di apprendimento, l'ACh elevata segnala che l'ambiente è strutturalmente stabile ma gli stimoli sono variabili, e quindi che occorre integrare più prove prima di aggiornare le rappresentazioni (basso tasso di apprendimento effettivo); la noradrenalina elevata segnala che la struttura stessa dell'ambiente è cambiata, e quindi che occorre discard le vecchie rappresentazioni e ricominciare da zero (alto tasso di apprendimento effettivo).

Il sistema noradrenergico e il gain globale. Il locus coeruleus (LC) è un piccolo nucleo del tronco cerebrale che contiene circa 15.000 neuroni nei primati, ma proietta noradrenalina (NA) su praticamente tutta la corteccia, il cerebello, il midollo spinale e molte strutture sottocorticali. Il profilo di risposta dei neuroni del LC è bimodale: in modalità *tonica* (firing rate basso ma continuo, ~1-3 Hz), la NA esercita un effetto di *arousal* globale che mantiene il cervello nello stato di veglia; in modalità *fasica* (brevi burst ad alta frequenza, > 10 Hz, in risposta a stimoli salenti), la NA produce un cambiamento di stato della rete corticale che facilita transizioni di comportamento e riorganizzazione delle rappresentazioni.

Aston-Jones e Cohen (2005) propongono nel loro influente modello *adaptive gain* che il LC ottimizza l'esplorazione vs. sfruttamento (*exploration-exploitation tradeoff*) modulando il guadagno delle funzioni di risposta delle cellule corticali. Ad alto guadagno noradrenergico, la curva risposta-input diventa più ripida: neuroni che preferiscono lo stimolo corrente firmano molto di più, mentre quelli che non lo preferiscono vengono silenziati — facilitando lo sfruttamento della rappresentazione corrente. A basso guadagno, la curva è più piatta: la differenza tra il neurone "vincitore" e gli altri si riduce, facilitando l'esplorazione di rappresentazioni alternative.

In termini di algoritmo di apprendimento, la NA modula effettivamente il *tasso di apprendimento* η in modo dinamico e adattivo: η cresce durante le transizioni ambientali (alta NA fasica) e si riduce durante la stabilità (bassa NA). Questo equivale a un *learning rate schedule* adattivo biologicamente implementato, analogamente — ma con meccanismi radicalmente diversi — agli ottimizzatori come Adam che adattano il tasso di apprendimento per ciascun parametro sulla base della storia dei gradienti.

Il sistema serotoninergico e la dimensione temporale dell'apprendimento. La serotonina (5-HT), rilasciata dai nuclei del rafe del tronco cerebrale, è spesso associata agli stati d'umore e all'ansia, ma ha un ruolo computazionale preciso nel contesto dell'apprendimento: modula il *fattore di sconto temporale* γ — la misura con cui il sistema valorizza le ricompense future rispetto a quelle immediate.

Dayan e Huys (2015) formalizzano questa intuizione in termini del TD-learning: la serotonina potrebbe implementare il complementare della dopamina nel framework actor-critic, codificando non l'errore di previsione della ricompensa immediata ma il valore attualizzato delle ricompense future su scale temporali più lunghe. Bassi livelli di 5-HT corrisponderebbero a basso γ — impazienza, svalutazione del futuro — mentre alti livelli corrisponderebbero a alto γ — capacità di tollerare le frustrazioni attuali per ricompense future maggiori. Questa ipotesi è coerente con il ruolo della 5-HT nella depressione (caratterizzata da deficit di pianificazione a lungo termine e da svalutazione del futuro) e con gli effetti degli inibitori della ricaptazione della serotonina (SSRI) nel ristabilire una prospettiva temporale più equa.

Interazioni tra sistemi neuromodulatori. I quattro sistemi non operano indipendentemente: le loro interazioni creano un sistema di meta-apprendimento adattivo multidimensionale. Eldar et al. (2016) mostrano che il rapporto tra NA e ACh determina lo stato di esplorazione/sfruttamento: alta NA / bassa ACh promuove l'esplorazione (alta incertezza, apprendimento rapido); bassa NA / alta ACh promuove lo sfruttamento (bassa incertezza, integrazione accurata delle prove). La dopamina, sovrapposta a questo sistema, segnala la direzione dell'aggiornamento (positivo o negativo). Il sistema serotoninergico scala il peso delle conseguenze temporalmente distanti. Insieme, questi quattro assi di modulazione definiscono uno spazio tetradimensionale di stati di apprendimento che il cervello esplora dinamicamente in risposta alle condizioni ambientali — un sistema di ottimizzazione del meta-apprendimento di una sofisticazione computazionale che i modelli di deep learning attuali non cominciano nemmeno ad abbozzare.

Calcolo dendritico non lineare: il neurone come rete profonda

La trattazione standard delle reti neurali artificiali modella ogni neurone come un'unità di integrazione lineare seguita da una non-linearità puntuale: il neurone somma le sue entrate ponderate e applica σ al risultato. Questa astrazione cattura la struttura di base della trasmissione sinaptica ma ignora la complessità computazionale intrinseca al singolo neurone biologico, che è — come hanno mostrato decenni di studi biofisici — essa stessa una piccola rete di elaborazione, non un semplice elemento di soglia.

Il neurone piramidale corticale ha una struttura dendritica elaborata: un albero dendritico apicale che può estendersi per oltre 1 mm nello strato molecolare, dendriti basali più corti che ricevono input dalle aree locali, e il compartimento iniziale dell'assone dove si integrano gli input somatici per generare potenziali d'azione. Crucialmente, i dendriti non sono conduttori passivi: contengono canali ionici voltaggio-dipendenti (Na^+ , Ca^{2+} , K^+) che

generano *potenziali dendritici attivi* locali che amplificano selettivamente gli input sinaptici forti e soppressi quelli deboli.

Nonlinearità dendritiche. Poirazi e Mel (2001) dimostrarono computazionalmente, e successivi esperimenti di patch-clamp in dendriti isolati confermarono, che i singoli rami dendritici si comportano come unità di integrazione sigmoide indipendenti: la risposta di un ramo dendritico al suo input sinaptico è una funzione sigmoide della somma degli input, con saturazione per input forti. L'intero neurone piramidale si comporta dunque come un sistema a *due strati*: nel primo strato, i singoli rami dendritici integrano i loro input con una non-linearità sigmoide locale; nel secondo strato, il soma integra le correnti dendritiche prodotte da tutti i rami con una seconda soglia di generazione del potenziale d'azione. Formalmente:

$$\text{output} = g(\sum_i w_i^{\text{soma}} \cdot \sigma(\sum w_i^{\text{dend}} \cdot x))$$

dove w_i^{dend} sono i pesi sinaptici sul ramo i , σ è la non-linearità sigmoide del ramo, w_i^{soma} è il peso dell'integrazione somatica del ramo i , e g è la funzione di soglia somatica. Questa struttura è quella di una rete neurale a due strati — ogni neurone è esso stesso una piccola rete profonda con capacità computazionale molto superiore a quella di un semplice perceptrone.

Poirazi e Mel stimarono che un neurone piramidale corticale con 10 rami dendritici attivi ha una capacità di classificazione lineare equivalente a circa 500 unità sigmoidali nella struttura point-neuron — una sottostima, perché esclude le interazioni non lineari tra rami.

Dendritic spikes. London e Häusser (2005) documentarono sperimentalmente la generazione di *dendritic spikes* — potenziali d'azione che si originano nel dendrite, non nel soma — nelle cellule piramidali di CA1 e di corteccia. I dendritic spikes si generano quando l'input sinaptico a un segmento dendritico supera una soglia critica di NMDA-LTP, producendo un burst locale di calcio che si propaga verso il soma dove può depolarizzare sufficientemente la membrana per generare un potenziale d'azione. La caratteristica cruciale è che i dendritic spikes possono propagarsi in modo decrementale lungo il dendrite, con una distanza di decadimento che dipende dalla distribuzione dei canali attivi locali, creando una forma di *nonlinear gating*: input distribuiti su rami diversi producono risposte sub-lineari (semplice somma delle correnti attenuate), ma input concentrati su un singolo ramo producono una risposta supralineare (dendritic spike che amplifica la depolarizzazione locale).

Implicazioni per l'attribuzione del credito. Il calcolo dendritico non lineare è direttamente rilevante per il problema dell'attribuzione del credito. Larkum et al. (2009) propongono che la coincidenza tra input basale (feedforward, dagli strati sottostanti) e input apicale (feedback, dagli strati superiori) sia la condizione necessaria per la generazione di dendritic spikes e la conseguente produzione di burst di potenziali d'azione al soma. La non-linearità dendritica implementa dunque un operatore di *coincidence detection* spazialmente preciso: solo le sinapsi basali che sono attive contemporaneamente all'input apicale ricevono la "firma" del burst (calcio locale, LTP) che segnala la loro partecipazione a una previsione confermata o a un errore rilevato. Sinapsi basali attive in assenza di coincidenza apicale non

ricevono questo segnale — la loro attivazione non è "certificata" come rilevante per la computazione corrente.

Questa proprietà trasforma il singolo neurone in un detector di coincidenza gerarchico: non soltanto corrispondenza temporale tra spike pre- e post-sinaptico (come nella STDP), ma corrispondenza spazio-temporale tra due stream informativi distinti — feedforward e feedback — che corrispondono rispettivamente all'evidenza sensoriale bottom-up e alla previsione top-down. La non-linearità dendritica è dunque il substrato biofisico che permette la separazione funzionale tra segnali di attivazione (input basale) e segnali di errore/conferma (input apicale) che il modello burstprop postula a livello teorico.

Reservoir computing e liquid state machines: apprendimento senza modificare i pesi profondi

Un paradigma computazionale radicalmente diverso da quelli discussi finora è il *reservoir computing*, sviluppato indipendentemente da Maass et al. (2002) nel framework delle *liquid state machines* (LSM) e da Jaeger (2001) nelle *echo state networks* (ESN). L'idea fondamentale è che il problema dell'attribuzione del credito negli strati profondi possa essere eluso completamente se si separano le reti in due componenti: un *reservoir* di connessioni fisse e casuali (che non vengono mai modificate dall'apprendimento) e un *readout* lineare addestrabile che mappa lo stato del reservoir sull'uscita desiderata.

Formalismo. Il reservoir è una rete ricorrente con N unità:

$$\mathbf{x}^{**}(t+1) = \sigma(\mathbf{W}_{\text{res}}^{**} \mathbf{x}^{**}(t) + \mathbf{W}_{\text{in}}^{**} \mathbf{u}^{**}(t) + \mathbf{b}^{**})$$

dove $\mathbf{W}_{\text{res}} \in \mathbb{R}^{(N \times N)}$ è la matrice dei pesi del reservoir (casuale, fissa), $\mathbf{W}_{\text{in}} \in \mathbb{R}^{(N \times d)}$ è la matrice di input (casuale, fissa), e $\mathbf{u}(t) \in \mathbb{R}^d$ è l'input al tempo t . Il readout è semplicemente:

$$\hat{\mathbf{y}}^{**}(t) = \mathbf{W}_{\text{out}}^{**} \mathbf{x}^{**}(t)$$

dove $\mathbf{W}_{\text{out}} \in \mathbb{R}^{(m \times N)}$ è l'unica matrice apprendibile, tipicamente ottimizzata per regressione lineare (ridge regression):

$$\mathbf{W}_{\text{out}}^{**} = \mathbf{Y}^{**} \mathbf{X}^{**T} (\mathbf{X}^{**} \mathbf{X}^{**T} + \lambda \mathbf{I}^{**})^{-1}$$

dove $\mathbf{X} \in \mathbb{R}^{(N \times T)}$ è la matrice degli stati del reservoir raccolti durante il training e $\mathbf{Y} \in \mathbb{R}^{(m \times T)}$ è la matrice dei target.

Echo state property. Il reservoir funziona correttamente se soddisfa la cosiddetta *echo state property*: lo stato del reservoir deve essere una funzione della storia dell'input, non delle condizioni iniziali. Condizione sufficiente è che il raggio spettrale $\rho(\mathbf{W}_{\text{res}}) < 1$ (in pratica

si usa $\rho \approx 0.8\text{--}0.99$). Con questa condizione, il reservoir ha una *fading memory*: la risposta a input lontani nel passato decade esponenzialmente, mentre i pattern recenti sono fedelmente rappresentati nello stato corrente. Il reservoir funge dunque da *memoria a breve termine non lineare*: proietta l'input recente in uno spazio ad alta dimensione che rende linearmente separabili pattern che nell'input originale erano non separabili.

Plausibilità biologica. Il reservoir computing è biologicamente plausibile in modo eccezionale rispetto agli altri paradigmi discussi, perché elimina completamente il requisito di propagare gradienti attraverso la rete: l'apprendimento avviene soltanto al readout, mediante un semplice aggiornamento lineare che può essere implementato da una regola di Hebb supervisionata con target locale. Maass et al. (2002) hanno argomentato che la corteccia cerebrale potrebbe operare come una collezione di reservoir — *liquid state machines* — con connettività ricorrente random (casualità che emerge dalla variabilità biologica nelle efficacie sinaptiche) e con aree "readout" (tipicamente la corteccia prefrontale o le aree motorie) che apprendono a leggere lo stato corrente dei circuiti sensori-motori per guidare il comportamento.

Jaeger e Haas (2004) mostrarono che reti ESN con readout addestrato possono generare segnali periodici complessi di grande precisione temporale — compiti che le RNN addestrate con BPTT trovano difficili per i problemi del vanishing gradient — suggerendo che la separazione tra reservoir fisso e readout apprendibile potrebbe essere una strategia efficace per apprendere dipendenze temporali a lungo raggio senza il costo computazionale della backpropagation through time.

Limitazioni. Il reservoir computing ha una limitazione fondamentale: la capacità espressiva del readout è lineare, per cui il sistema può apprendere soltanto trasformazioni lineari dello spazio di stati del reservoir. Se il compito richiede una mappatura non lineare delle caratteristiche estratte dal reservoir, il readout lineare non è sufficiente. Estensioni con readout non lineare (ESN profonde, deep reservoir networks) affrontano questo problema a costo di reintrodurre parzialmente il problema dell'attribuzione del credito. Maass et al. hanno proposto che nel cervello il readout corticale sia implementato da connessioni apprendibili (con regole hebbiane locali supervisionate dal segnale di rinforzo) dalla corteccia sensoriale alla corteccia motoria o prefrontale, mantenendo il reservoir corticale (le connessioni ricorrenti intra-area e inter-area) essenzialmente invariato nel corso dell'apprendimento di singoli compiti.

Periodi critici dello sviluppo: apertura e chiusura delle finestre plastiche

Un aspetto dell'apprendimento biologico che non ha alcun corrispondente nelle reti artificiali standard è la struttura temporale dello sviluppo: il fatto che il cervello abbia periodi critici (*critical periods*, CP) — finestre temporali durante le quali la plasticità è massima e le strutture corticali si formano irreversibilmente in risposta all'esperienza sensoriale — seguiti da periodi di ridotta plasticità in cui le strutture consolidate sono resistenti alla modificazione.

Il periodo critico visivo. Hubel e Wiesel (1970) e Wiesel e Hubel (1963) documentarono il primo periodo critico in V1 del gatto: se un occhio viene suturato nei primi mesi di vita, le

colonne di dominanza oculare si riorganizzano a favore dell'occhio aperto, e la deprivazione prolungata produce ambliopia permanente che non si risolve anche se la sutura viene rimossa in età adulta. Questo fenomeno — la *ocular dominance plasticity* — è il prototipo dei CP. La finestra critica si apre intorno alla quarta settimana di vita (nel gatto), raggiunge un picco intorno alla sesta-ottava settimana, e si chiude entro il terzo-quarto mese.

I meccanismi molecolari dell'apertura e chiusura dei CP sono stati chiariti nell'arco degli ultimi vent'anni. L'apertura è mediata dall'BDNF (Brain-Derived Neurotrophic Factor) che accelera la maturazione degli interneuroni PV+ — gli stessi che implementano l'inibizione laterale e il controllo del guadagno discussi sopra. Strangely (2004) e Fagiolini et al. (2004) mostrarono che è la maturazione degli interneuroni PV+ a *attivare* il periodo critico, non semplicemente un aumento generico di attività. In assenza di interneuroni PV+ funzionali (topi knockout per la parvalbumin), il periodo critico non si apre: la rete manca del bilanciamento eccitazione/inibizione necessario per la plasticità strutturale.

La chiusura dei CP è mediata dall'accumulo di *Perineuronal Nets* (PNN) — strutture di matrice extracellulare che avvolgono fisicamente i neuroni PV+ riducendo la loro mobilità sinaptica e stabilizzando i circuiti. La degradazione delle PNN mediante condroitinasi (un enzima batterico) in animali adulti riapre parzialmente il periodo critico, ripristinando la plasticità della dominanza oculare — una dimostrazione che le PNN non sono solo un epifenomeno dello sviluppo ma un meccanismo attivo di consolidazione strutturale.

Implicazioni computazionali. I periodi critici implementano biologicamente una forma di *curriculum learning* strutturato: le diverse aree corticali hanno periodi critici con timing diverso, organizzati in modo tale che le aree di ordine inferiore (V1, A1) si strutturano prima, quando l'input sensoriale è semplice (attività spontanea retinica o cocleare prenatale), e le aree di ordine superiore si strutturano più tardi, quando possono sfruttare le rappresentazioni già consolidate degli strati inferiori come input. Questa organizzazione sequenziale ricapitola il curriculum dell'addestramento nei moderni sistemi deep learning, in cui i modelli pre-addestrati su compiti semplici vengono fine-tuned su compiti più complessi — ma è implementata attraverso un programma geneticamente controllato di apertura e chiusura delle finestre plastiche, non da una scelta esplicita dell'addestratore.

Zador (2019) ha argomentato che la struttura dei periodi critici è essa stessa il prodotto della selezione evolutiva: l'evoluzione ha ottimizzato non soltanto la struttura del connettoma adulto, ma l'intera traiettoria di sviluppo — il programma di apertura e chiusura delle finestre plastiche, le sequenze di maturazione delle aree corticali, la progressione dalla dipendenza dall'attività spontanea alla dipendenza dall'esperienza. In questa prospettiva, l'ontogenesi è il mezzo attraverso cui l'evoluzione garantisce la robustezza dello sviluppo: il connettoma adulto emerge da un processo guidato di apprendimento biologico, non da una singola specifica genetica fissa.

Il sistema artificiale più vicino ai CP è la famiglia dei metodi di *progressive training*: reti in cui i parametri degli strati inferiori vengono congelati dopo un certo numero di passi di addestramento, permettendo agli strati superiori di adattarsi su rappresentazioni stabili degli strati inferiori. Curriculum learning (Bengio et al., 2009) generalizza questa idea, mostrando che l'ordine di presentazione degli esempi durante l'addestramento influisce sulla qualità

finale della rete — esattamente come la sequenza temporale degli input durante i periodi critici determina la struttura finale delle rappresentazioni corticali.

Apprendimento contrastivo e auto-supervisione: convergenza verso la biologia

Negli ultimi anni, il paradigma dell'apprendimento auto-supervisionato (*self-supervised learning*) ha rivoluzionato il pre-training dei modelli di linguaggio e visione, avvicinando il machine learning alle modalità di apprendimento biologico in modo più diretto di quanto si potesse anticipare.

I metodi di *contrastive self-supervised learning* — SimCLR (Chen et al., 2020), MoCo (He et al., 2020), BYOL (Grill et al., 2020) — addestrano una rete a produrre rappresentazioni simili per due augmented views dello stesso dato (coppia positiva) e dissimili per dati diversi (coppia negativa). La funzione di perdita contrastiva *InfoNCE* è:

$$\mathcal{L} = -\mathbb{E}[\log (e^{(z \cdot z^+ / \tau)}) / (e^{(z \cdot z^+ / \tau)} + \sum e^{(z \cdot z^- / \tau)})]$$

dove z , z^+ sono le rappresentazioni della coppia positiva e z^- sono le rappresentazioni delle coppie negative, e τ è una temperatura. Minimizzare questa perdita è equivalente a massimizzare l'informazione mutua tra le due viste dello stesso dato, implementando il principio di massimizzazione dell'informazione mutua di Linsker (1988) — uno dei framework più antichi per l'apprendimento non supervisionato biologicamente motivato.

Il collegamento con la biologia è diretto: il cervello non riceve etichette supervisionate, ma è esposto ripetutamente agli stessi oggetti in condizioni diverse (diversa illuminazione, diverso punto di vista, diverso contesto sonoro) e deve costruire rappresentazioni invarianti a queste trasformazioni. L'apprendimento contrastivo implementa esattamente questo principio: l'invarianza alle trasformazioni emerge come conseguenza dell'ottimizzazione di un obiettivo di coerenza temporale, senza supervisione esplicita. Simoncelli e Olshausen (2001) avevano già proposto che le cellule semplici di V1 possano essere viste come il risultato dell'ottimizzazione di un obiettivo di massimizzazione dell'informazione mutua tra input e rappresentazione sotto un vincolo di sparsità — una versione non contrastiva dello stesso principio.

LeCun (2022), nella sua proposta di *Joint Embedding Predictive Architecture* (JEPA), ha arguito che l'architettura dell'apprendimento auto-supervisionato più vicina al funzionamento biologico non è quella contrastiva (che richiede coppie negative, potenzialmente instabile) ma quella *predittiva*: il sistema impara a predire la rappresentazione di una parte del dato da un'altra parte, nello spazio delle rappresentazioni (non nello spazio dei pixel). Questa architettura — un'estensione del predictive coding al dominio dell'auto-supervisione — elimina la necessità di confronto esplicito con campioni negativi e implementa un principio di coerenza interna analogo a quello del predictive coding gerarchico, ma applicato all'asse spaziale piuttosto che a quello temporale predittivo.

Il problema della generalizzazione: distribuzione di test vs. distribuzione di training

Una delle differenze più profonde tra apprendimento biologico e apprendimento con backpropagation è la robustezza alla distribuzione: i sistemi biologici generalizzano con straordinaria efficacia a contesti mai incontrati prima, mentre le reti addestrate con backpropagation mostrano un degrado spesso catastrofico quando la distribuzione di test differisce dalla distribuzione di training — il problema del *distribution shift* o *out-of-distribution generalization*.

Illusioni ottiche e robustezza. Un esempio notevole è la risposta dei sistemi di visione artificiale agli *adversarial examples* (Szegedy et al., 2014): piccole perturbazioni impercettibili delle immagini di input, ottimizzate per massimizzare l'errore di classificazione, ingannano le CNN con straordinaria efficacia (tasso di errore > 99%) mentre sono completamente invisibili agli umani. Questo fenomeno, noto come *adversarial vulnerability*, indica che le CNN hanno appreso rappresentazioni statisticamente correlate con le etichette di training ma non corrispondenti alle caratteristiche causalmente rilevanti per il compito. Il cervello umano — che utilizza meccanismi di apprendimento molto più lenti, con efficienza campionaria enormemente superiore — sembra aver sviluppato rappresentazioni più causalmente robuste, che resistono alle perturbazioni superficiali degli input.

Causal representation learning. Schölkopf et al. (2021) propongono che la chiave per la robustezza al distribution shift sia l'apprendimento di *rappresentazioni causali*: rappresentazioni che riflettono le relazioni causa-effetto tra variabili del mondo, piuttosto che le sole correlazioni statistiche. La backpropagation è intrinsecamente adirezionale in questo senso: minimizza la perdita sulla distribuzione di training senza alcun bias verso la struttura causale del processo generativo dei dati. Il cervello, invece, sembra operare su un modello del mondo che incorpora struttura causale — evidenziato dalla capacità di controfattuale (*counterfactual reasoning*): gli umani possono valutare "cosa sarebbe successo se X fosse stato diverso" anche in assenza di esperienza diretta di quella situazione.

La connessione con il predictive coding è diretta: un modello generativo del tipo $p(\mathbf{x}|\mathbf{z})$ con variabili latenti strutturate causalmente è intrinsecamente più robusto a distribution shifts che non modificano la struttura causale, perché le variabili latenti rimangono invarianti anche quando la distribuzione delle osservazioni cambia. L'apprendimento del modello generativo mediante minimizzazione della free energy — che il predictive coding implementa biologicamente — è dunque non soltanto biologicamente plausibile ma anche computazionalmente motivato per la generalizzazione robusta.

Parte VI — Convergenze, NeuroAI e la domanda che rimane

Il campo emergente della NeuroAI

Il termine *NeuroAI* è entrato nel lessico scientifico circa nel 2021, quando Richards et al. pubblicarono un articolo manifesto su *Nature Neuroscience* intitolato *A deep learning framework for neuroscience*, proponendo che le reti neurali profonde debbano essere considerate non soltanto strumenti di ingegneria ma *modelli computazionali* del cervello, soggetti alle stesse metodologie di validazione delle teorie scientifiche. Questo programma ha una doppia ambizione: da un lato, usare il deep learning come telescopio per guardare il cervello con una risoluzione senza precedenti; dall'altro, usare la biologia per costruire sistemi artificiali più robusti, efficienti e generalizzabili.

La NeuroAI non è però un campo omogeneo. Si possono distinguere almeno tre correnti intellettuali con premesse e metodi diversi.

La prima corrente, che potremmo chiamare *neuroscienza inversa per mezzo del deep learning*, usa modelli artificiali addestrati su compiti definiti per generare previsioni sull'organizzazione funzionale delle aree corticali. Il paradigma metodologico è quello di Yamins e DiCarlo: si addestra una CNN su classificazione visiva, si confrontano le sue rappresentazioni interne con le risposte neuronali, e le discrepanze sistematiche guidano lo sviluppo di architetture migliori — un ciclo iterativo tra modello e dati sperimentali. Il successo di questo approccio nel predire le risposte di IT della scimmia ha già influenzato il disegno di esperimenti neurobiologici: si cerca sperimentalmente ciò che i modelli computazionali predic-ono, invece di esplorare ciecamente lo spazio delle risposte possibili.

La seconda corrente, che potremmo chiamare *ingegneria biologicamente ispirata*, usa i meccanismi biologici come sorgente di idee architettureali per superare le limitazioni dei sistemi attuali. I risultati più fecondi di questa corrente sono stati l'attenzione — ispirata ai meccanismi di selezione selettiva dell'informazione in corteccia — i meccanismi di memoria esterna (Neural Turing Machine di Graves et al., 2014; Differentiable Neural Computer di Graves et al., 2016) — ispirati alla complementarità ippocampo/neocorteccia — e i metodi di continual learning descritti nella Parte IV.

La terza corrente, forse la più ambiziosa intellettualmente, tenta di costruire una teoria unificata dell'intelligenza che trascenda la distinzione tra biologia e ingegneria. Bengio et al. (2021) propongono la *GFlowNet* come meccanismo di inferenza generativa che rispecchia alcune proprietà dell'apprendimento biologico: al posto di cercare il singolo punto di massima verosimiglianza (come fa la discesa del gradiente), la GFlowNet campiona dalla distribuzione completa delle soluzioni proporzionalmente alla loro verosimiglianza, producendo diverse ipotesi plausibili — un comportamento più simile al ragionamento abduttivo umano che all'ottimizzazione deterministica. Questa proposta è ancora allo stadio di esplorazione, ma rappresenta una direzione in cui la distinzione tra "biologia" e "ingegneria" tende a dissolversi in favore di una teoria generale dell'apprendimento adattivo.

Il connettoma come archivio evolutivo dell'ottimizzazione

Una delle intuizioni più profonde emerse dalla neurobiologia strutturale degli ultimi anni è che il connettoma — la mappa completa delle connessioni sinaptiche di un organismo — non è semplicemente il risultato dell'apprendimento individuale, ma un archivio accumulato di milioni di anni di ottimizzazione evolutiva. Ogni aspetto della struttura sinaptica, dalla distribuzione dei tipi di recettori ai pattern di connettività delle diverse classi di interneuroni, porta l'impronta di pressioni selettive che hanno agito in ambienti fisici e sociali specifici.

Questa prospettiva ha implicazioni profonde per l'interpretazione dei meccanismi di apprendimento biologico. Se la maggior parte dell'informazione adattiva è codificata nel connettoma innato, allora i meccanismi di plasticità sinaptica come LTP, LTD e STDP svolgono un ruolo relativamente modesto nella costruzione delle competenze cognitive — sono meccanismi di *affinamento locale* su una struttura già fortemente vincolata, non di costruzione dal basso. Questa visione è in tensione con la narrativa dominante del deep learning, in cui tutta l'intelligenza emerge dall'ottimizzazione tabula rasa mediante backpropagation.

La questione può essere formalizzata nei termini della distinzione tra *prior* e *likelihood* nell'inferenza bayesiana. Il connettoma innato costituisce un prior fortissimo — codificato biologicamente — sullo spazio delle rappresentazioni utili. La backpropagation (o i suoi equivalenti biologici) opera come aggiornamento della likelihood in risposta all'esperienza individuale, spostando la distribuzione a posteriori dal prior verso la soluzione compatibile con i dati. In un sistema in cui il prior è molto informativo (connettoma evolutivamente ottimizzato), il contributo della likelihood è relativamente piccolo — pochi passi di aggiornamento sono sufficienti per produrre comportamenti adattativi, perché il sistema parte già vicino alla soluzione. È questo, probabilmente, il meccanismo alla base dell'apprendimento few-shot biologico: non un algoritmo di apprendimento straordinariamente efficiente, ma un prior straordinariamente preciso.

Il paradosso dell'uovo e della gallina. C'è tuttavia una circolarità in questo argomento: il prior evolutivamente ottimizzato è a sua volta il risultato di un processo di "apprendimento" su scala evolutiva — la selezione naturale che ha selezionato le strutture connettomali più adatte agli ambienti ancestrali. Questo processo non usa la backpropagation (né potrebbe farlo su scale temporali evolutive), ma è funzionalmente equivalente a una forma di ottimizzazione che minimizza la mortalità e massimizza la riproduzione su una distribuzione di ambienti ancestrali. La domanda "come il cervello apprende?" si sdoppia dunque in due domande distinte: "come il cervello individuale si adatta nell'arco della vita?" e "come la struttura del cervello è stata ottimizzata dall'evoluzione?". La backpropagation (e i suoi equivalenti biologici) è rilevante per la prima; la teoria della selezione naturale, la genetica dello sviluppo e l'evo-devo sono rilevanti per la seconda. La separazione analitica tra questi due livelli è necessaria per evitare confusioni concettuali, sebbene i due processi interagiscano profondamente.

Reti di spiking: il calcolo come computazione d'impulso

Tutte le reti neurali artificiali discusse in questa monografia usano attivazioni continue — numeri reali che rappresentano tassi di firing medi. Il cervello biologico, invece, comunica mediante *spikes* — potenziali d'azione discreti e binari. Questa differenza non è soltanto di dettaglio implementazionale: le reti di spiking hanno proprietà computazionali e di apprendimento radicalmente diverse dalle reti rate-based.

Il modello leaky integrate-and-fire (LIF). Il modello di neurone spiking più semplice che cattura le proprietà essenziali è il *leaky integrate-and-fire*.

$$\tau_m \frac{du}{dt} = -(u - u_{\text{rest}}) + R_m I(t)$$

dove u è il potenziale di membrana, $\tau_m = R_m C_m$ è la costante di tempo di membrana, R_m è la resistenza di membrana, C_m è la capacitance, e $I(t)$ è la corrente totale in ingresso (sinaptica + eventuale corrente di stimolazione). Quando u raggiunge la soglia ϑ , il neurone genera un spike e viene resettato a $u_{\text{reset}} < \vartheta$. La corrente sinaptica prodotta da uno spike del neurone pre-sinaptico j al tempo t^s è modellata come un esponenziale decrescente:

$$I_j(t) = \sum w_j \cdot \exp(-(t - t^s)/\tau_{\text{syn}}) \cdot \mathbb{1}[t > t^s]$$

Spike-based coding e temporale coding. Le reti di spiking permettono due modalità di codifica dell'informazione qualitativamente diverse. La *rate coding* usa il tasso di firing (numero di spikes nell'unità di tempo) come variabile di codifica — riducendo il neurone spiking a un'approssimazione rumorosa del neurone a tasso continuo. La *temporal coding* usa il tempo preciso degli spike come portatore di informazione: la latenza del primo spike dopo la presentazione di uno stimolo (time-to-first-spike coding), la fase dello spike rispetto a un'oscillazione di riferimento (phase coding), o pattern specifici di spike-timing (polychronous networks, Izhikevich 2006).

Le reti spiking con temporal coding hanno proprietà computazionali interessanti: Thorpe et al. (2001) hanno mostrato che il cervello umano può riconoscere scene visive complesse entro 150-200 ms — tempo sufficiente per soltanto 2-3 cicli sinaptici se il riconoscimento dipende dal solo tasso di firing (che richiede finestre di integrazione di almeno 50-100 ms). Se invece l'informazione è codificata nella latenza del primo spike, il riconoscimento può avvenire in un singolo ciclo sinaptico, coerente con i tempi osservati. Questo suggerisce che almeno una parte dell'elaborazione visiva rapida non usa rate coding ma temporal coding basato sul primo spike.

Backpropagation nelle reti spiking. Addestrare reti spiking con backpropagation è tecnicamente difficile a causa della non-differenziabilità della funzione di spike: la variabile di output è binaria (spike/no spike), e il suo gradiente rispetto al potenziale di membrana è zero quasi ovunque e indefinito alla soglia. Diverse soluzioni sono state proposte:

La *surrogate gradient method* (Neftci et al., 2019; Bellec et al., 2020) sostituisce la funzione di spike non-differenziabile con un *surrogate gradient* differenziabile durante il backward pass — tipicamente una sigmoide o una gaussiana centrata alla soglia — pur mantenendo la funzione di spike discreta durante il forward pass. Questo trucco, analogo al *straight-through estimator* per

variabili discrete nelle reti di quantizzazione, permette l'addestramento di reti spiking profonde con prestazioni competitive su benchmark standard.

Il metodo dei surrogate gradients è biologicamente rilevante perché mostra che la discontinuità degli spikes non è un ostacolo insuperabile per l'apprendimento basato su gradienti: qualunque approssimazione del gradiente che punti nella direzione corretta è sufficiente per la convergenza, e la non-linearità del threshold può essere "ammorbidita" localmente dalla distribuzione stocastica dei tempi di spike.

SpikeProp e SuperSpike. Bohte et al. (2000) proposero *SpikeProp*, la prima implementazione della backpropagation per reti spiking con temporal coding: il segnale di errore è la differenza tra il tempo del spike effettivo e il tempo del spike desiderato, e si propaga attraverso una regola della catena che include le derivate dei tempi di spike rispetto ai pesi. Zenke e Ganguli (2018) hanno sviluppato *SuperSpike*, una regola di apprendimento supervisionata per reti spiking che risolve il problema del gradiente non-differenziabile usando un'approssimazione di secondo ordine del gradiente della perdita:

$$\Delta w_i = -\eta \cdot [\partial \mathcal{L} / \partial u_i]_{\text{surrogate}} \cdot \int_{-\infty}^t \alpha(t-s) x(s) ds$$

dove $\alpha(t)$ è un filtro esponenziale che interpola la traccia della pre-sinapsi, e $[\partial \mathcal{L} / \partial u_i]_{\text{surrogate}}$ è il surrogate gradient. Questa regola è biologicamente plausibile (locale, dipende dalla traccia pre-sinaptica e dall'errore post-sinaptico) e produce reti spiking con prestazioni competitive con le reti rate-based su compiti di classificazione temporale.

Verso una teoria del campionamento neurale

Una prospettiva teorica di grande originalità è quella che interpreta l'attività stocastica delle reti neurali non come rumore, ma come un meccanismo computazionale deliberato di *campionamento dalla distribuzione a posteriori*. Questa idea, sviluppata indipendentemente da Hoyer e Hyvärinen (2003), Fiser et al. (2010) e Orbán et al. (2016), propone che la variabilità del firing neuronale codifichi l'incertezza della rappresentazione: invece di rappresentare la stima di massima a posteriori di una variabile latente (come farebbe un neurone deterministico), ogni neurone produce campioni dalla sua distribuzione a posteriori marginale.

Il framework. Consideriamo un modello probabilistico in cui lo stato nascosto $\mathbf{z}$ ha una distribuzione a priori $p(\mathbf{z})$ e la likelihood è $p(\mathbf{x}|\mathbf{z})$. La distribuzione a posteriori $p(\mathbf{z}|\mathbf{x})$ è la distribuzione "ottimale" per le rappresentazioni date le osservazioni. Il framework del campionamento neurale propone che la traiettoria temporale dell'attività neuronale $\{\mathbf{z}(t)\}$ sia un processo di Markov che campiona asintoticamente dalla distribuzione $p(\mathbf{z}|\mathbf{x})$. Formalmente:

$$p_{\text{stationary}}(\mathbf{z}) = p(\mathbf{z}|\mathbf{x})$$

Una condizione sufficiente è che le equazioni del moto soddisfino il *detailed balance* rispetto a $p(\mathbf{z}|\mathbf{x})$:

$$p(\mathbf{z}^{**}|\mathbf{x}^{**}) q(\mathbf{z}^{**} \rightarrow \mathbf{z}^{**'}) = p(\mathbf{z}^{**'}|\mathbf{x}^{**}) q(\mathbf{z}^{**'} \rightarrow \mathbf{z}^{**})$$

dove $q(\mathbf{z} \rightarrow \mathbf{z}')$ è la probabilità di transizione delle dinamiche neurali. Reti con architettura e sinergie opportune possono soddisfare questa condizione, producendo una catena di Markov ergodica con stazionaria uguale alla posteriori desiderata.

Connessione con il predictive coding. Nel framework del campionamento neurale, il predictive coding non è soltanto un meccanismo per calcolare la stima MAP della rappresentazione (il minimo dell'energia libera), ma il *campo di deriva* di un processo stocastico che campiona dalla distribuzione a posteriori. Il rumore sinaptico — tipicamente considerato un fattore da minimizzare — diventa funzionale: la varianza del firing del neurone codifica la varianza della sua distribuzione marginale a posteriori, e il campionamento stocastico garantisce la copertura dell'intera distribuzione, non solo del suo modo.

Orbán et al. (2016) testarono questa predizione in V1 di scimmia: la variabilità dello spike-count in risposta agli stimoli naturali dovrebbe essere proporzionale all'incertezza delle rappresentazioni (alta per stimoli ambigui, bassa per stimoli chiari). I dati confermano questa predizione quantitativamente, con la variabilità del firing che scala con la misura di incertezza bayesiana in modo statisticamente consistente con il campionamento dalla posteriori.

Implicazioni per l'apprendimento. Il framework del campionamento neurale ha implicazioni dirette per l'apprendimento: se le reti neurali campionano dalla posteriori, allora l'aggiornamento dei pesi dovrebbe massimizzare la verosimiglianza dei dati sotto il modello generativo, non minimizzare un errore quadratico. Questo equivale all'apprendimento nell'algoritmo EM (Expectation-Maximization): il passo E calcola la distribuzione a posteriori sui dati nascosti (implementato dal campionamento neurale), e il passo M aggiorna i parametri per massimizzare la log-verosimiglianza attesa. L'algoritmo wake-sleep di Hinton et al. (1995) è una versione approssimata di EM che condivide questa struttura a due fasi — il collegamento con il ciclo sonno/veglia biologico assume dunque, in questa prospettiva, un significato computazionale preciso: la fase "veglia" corrisponde al passo E (campionamento della posteriori data l'evidenza), mentre la fase "sonno" corrisponde al passo M (aggiornamento del modello generativo sulla base dei campioni generati internamente).

Ricorrenza e feedback: l'architettura dimenticata

Un aspetto dell'architettura corticale che le reti neurali artificiali standard trascurano sistematicamente è la pervasività delle connessioni ricorrenti e di feedback. Nel cervello, le connessioni feedback (da aree superiori a inferiori) sono quantitativamente superiori alle connessioni feedforward (Felleman e Van Essen, 1991): per ogni fibra feedforward che connette V1 a V2, ci sono tipicamente 10 fibre feedback dalla direzione opposta. Questo

rapporto 10:1 — completamente invertito rispetto all'architettura di una CNN standard — non è un caso anatomico privo di significato: riflette il fatto che la corteccia è prima di tutto un sistema predittivo che usa la propria attività passata per generare previsioni, e soltanto secondariamente un sistema di classificazione feedforward.

La ricorrenza intra-areale — le connessioni orizzontali tra neuroni della stessa area — è altrettanto abbondante. In V1, circa il 70% del volume di dendriti di ciascuna cellula piramidale riceve input da altre cellule della stessa area, con un range spaziale di diversi mm — sufficiente per connettere neuroni con campi recettivi che si sovrappongono solo parzialmente. Queste connessioni orizzontali implementano computazioni di *contextual modulation*: la risposta di un neurone V1 a uno stimolo nel suo campo recettivo è modulata dall'attività di neuroni con campi recettivi adiacenti — un meccanismo che implementa la percezione della struttura di sfondo e del contorno globale, computazioni che le CNN feedforward non possono replicare con un singolo livello.

Reti ricorrenti e temporal depth. Dal punto di vista del deep learning, la ricorrenza corticale equivale ad aggiungere una profondità temporale alla rete: ogni passo temporale di elaborazione ricorrente corrisponde a un ulteriore strato di processamento. Una rete corticale con k strati gerarchici e T cicli temporali di elaborazione ricorrente è computazionalmente equivalente a una rete feedforward con $k \cdot T$ strati effettivi — una profondità molto maggiore di quanto le sole k aree anatomiche suggerirebbero. Questa equivalenza è sfruttata esplicitamente nei modelli CORnet (Kubilius et al., 2019), in cui le connessioni ricorrenti intra-area permettono di aumentare la profondità computazionale effettiva senza aumentare il numero di parametri.

Backpropagation nelle reti ricorrenti biologiche. Per addestrare reti ricorrenti con backpropagation, si usa BPTT — che richiede di memorizzare tutti gli stati intermedi per l'intera lunghezza della sequenza, con un costo in memoria proporzionale a T . Nelle reti biologiche, T può essere dell'ordine di secondi (per i cicli di elaborazione corticale) o di ore (per i cicli di consolidamento durante il sonno), rendendo la BPTT computazionalmente proibitiva anche in termini di requisiti di memoria biologica. Le soluzioni biologiche discusse nelle Parti II e III — tracce di eligibilità, plateau potentials nell'ippocampo, replay durante i SWR — sono tutte varianti di meccanismi che approssimano il gradiente temporale senza richiedere la memorizzazione esplicita di tutti gli stati passati, mantenendo soltanto un sottoinsieme compresso dell'informazione storica rilevante.

Il dualismo cognizione-percezione nel contesto dell'apprendimento

La domanda "il cervello usa la backpropagation?" assume un significato completamente diverso a seconda di quale sistema cognitivo si consideri. Il cervello non è un sistema uniforme: le proprietà di apprendimento della corteccia visiva primaria sono radicalmente diverse da quelle della corteccia prefrontale, e quelle dell'ippocampo sono diverse da entrambe. Confondere questi sistemi — come spesso accade nella letteratura del confronto CNN-cervello — porta a conclusioni valide per un sottosistema specifico ma erranee in senso generale.

Apprendimento percettivo vs. apprendimento concettuale. Nei sistemi sensoriali primari (V1, A1, S1), l'apprendimento è lento, richiede migliaia di prove, è altamente specifico per le caratteristiche dello stimolo, e le modifiche si stabilizzano durante il sonno — un profilo compatibile con un apprendimento di tipo hebbiano lentamente accumulato, mediato da modifiche sinaptiche graduali indotte dalla STDP e dall'attività correlata. Nei sistemi di ordine superiore (corteccia prefrontale, corteccia temporale anteriore, ippocampo), l'apprendimento è rapido (spesso one-shot), flessibile, dipendente dal contesto, e capace di generalizzazione immediata — un profilo più compatibile con il framework del meta-apprendimento o con meccanismi di retrieval associativo ad alta capacità.

Questa eterogeneità suggerisce che il cervello non utilizzi un singolo algoritmo di apprendimento, ma una gerarchia di algoritmi con proprietà complementari: i sistemi di ordine inferiore implementano qualcosa di più simile all'apprendimento hebbiano auto-supervisionato (addestrando lentamente rappresentazioni gerarchiche), mentre i sistemi di ordine superiore implementano qualcosa di più simile all'apprendimento per rinforzo con memoria episodica ad alta capacità (addestrando rapidamente l'uso adattivo di quelle rappresentazioni). La backpropagation — o i suoi equivalenti biologici — potrebbe essere rilevante per i sistemi di ordine inferiore, ma probabilmente non è il meccanismo dominante nei sistemi di ordine superiore dove il numero di prove è troppo limitato per la convergenza di un gradiente discendente.

L'orizzonte del problema: cosa sappiamo e cosa non sappiamo

Giunti alla conclusione di questa monografia, è possibile tracciare una mappa di ciò che sappiamo con ragionevole certezza, di ciò che è ragionevolmente ipotizzabile, e di ciò che rimane genuinamente ignoto.

Sappiamo con certezza che la backpropagation nella sua forma letterale — con simmetria dei pesi, blocco modulante, separazione netta training/inferenza e propagazione di gradienti in virgola mobile a precisione elevata — non è implementata nel substrato neuronale biologico. Ogni assunzione strutturale dell'algoritmo è in contraddizione con almeno una proprietà nota della fisiologia sinaptica o dell'anatomia corticale. Questa conclusione è robusta e non controversa nella comunità scientifica.

Sappiamo con ragionevole certezza che esistono meccanismi biologici — predictive coding, BTSP ippocampale, regola burstprop, apprendimento contrastivo mediato da circuiti oscillatori — che approssimano il gradiente della funzione di perdita con metodi locali e biologicamente plausibili. La qualità di questa approssimazione è teoricamente analizzata per architetture semplificate, e le previsioni quantitative di alcuni di questi meccanismi sono state confermate sperimentalmente.

È ragionevolmente ipotizzabile che il cervello utilizzi una combinazione di questi meccanismi in circuiti diversi e per compiti diversi, coordinati da segnali neuromodulatori che controllano adattativamente il tasso di apprendimento, la selettività della plasticità, e il bilanciamento tra esplorazione e consolidamento. L'evidenza per questa ipotesi è indiretta ma convergente da diverse direzioni sperimentali.

Non sappiamo se la somma di questi meccanismi sia sufficiente a spiegare la totalità dell'apprendimento biologico osservato, specialmente per le forme più complesse e flessibili di generalizzazione e ragionamento astratto che distinguono la cognizione dei primati. Non sappiamo se esista un meccanismo ancora non identificato — forse coinvolgente le cellule gliali, forse le oscillazioni di rete su scale temporali molto lunghe, forse dinamiche epigenetiche che modificano la struttura sinaptometabolica a lungo termine — che risolva il problema dell'attribuzione del credito in modo più completo di quanto i meccanismi attualmente noti possano fare.

Ciò che sappiamo con certezza, in ultima analisi, è che il problema dell'attribuzione del credito non è soltanto un problema algoritmico ma un problema fisico: qualunque sistema materiale che si adatti al proprio ambiente deve trovare un modo di "chiudere il cerchio" tra la valutazione dell'esito e la modifica dei meccanismi che lo hanno prodotto. In questa luce, la backpropagation è un modo elegante e computazionalmente efficiente di chiudere questo cerchio in un substrato digitale, e l'apprendimento biologico è un modo biologicamente parsimonioso di avvicinarsi alla stessa soluzione in un substrato materiale governato da leggi fisiche radicalmente diverse. Che due percorsi così diversi convergano verso soluzioni funzionalmente analoghe — e che questa convergenza si possa dimostrare formalmente — è forse il risultato più bello che l'incontro tra neuroscienze e machine learning ha prodotto sino ad oggi.

Appendici Tecniche

Appendice A — Dimostrazione del teorema di approssimazione universale

Il teorema di approssimazione universale (Cybenko, 1989; Hornik, Stinchcombe e White, 1989) è il fondamento teorico che giustifica l'uso delle reti neurali come modelli flessibili. Presentiamo la versione di Cybenko per reti a uno strato nascosto con attivazioni sigmoidei.

Enunciato. Sia $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ una funzione continua e sigmoide (nel senso che $\sigma(t) \rightarrow 1$ per $t \rightarrow +\infty$ e $\sigma(t) \rightarrow 0$ per $t \rightarrow -\infty$). Sia $I = [0,1]^n$ il cubo unitario n -dimensionale e $C(I)$ lo spazio delle funzioni continue su I con la norma del sup. Allora, per ogni $f \in C(I)$ e ogni $\varepsilon > 0$, esistono un intero N , vettori $\mathbf{w} \in \mathbb{R}^n$, bias $b \in \mathbb{R}$ e coefficienti $\alpha \in \mathbb{R}$ con $j = 1, \dots, N$, tali che la funzione

$$F(\mathbf{x}) = \sum_{j=1}^N \alpha_j \sigma(\mathbf{w}_j^T \mathbf{x} + b_j)$$

soddisfa $\|F - f\|_\infty < \varepsilon$.

Dimostrazione (schema). La prova si articola in tre passi.

Passo 1. Si mostra che l'insieme delle funzioni della forma $\Sigma \propto \sigma(\mathbf{w}^T \mathbf{x} + b)$ forma uno spazio vettoriale Σ chiuso rispetto alla convergenza uniforme su I . Questo è diretto per linearità di Σ .

Passo 2. Si dimostra per assurdo che Σ è denso in $C(I)$. Si suppone che Σ non sia denso; allora, per il teorema di Hahn-Banach nella forma funzionale, esiste una misura di Borel regolare μ su I , non identicamente nulla, tale che:

$$\int_{\mathbb{R}} \sigma(\mathbf{w}^T \mathbf{x} + b) d\mu(\mathbf{x}) = 0 \quad \text{per tutti } \mathbf{w} \in \mathbb{R}^n, b \in \mathbb{R}$$

Passo 3. Si mostra che questa condizione implica $\mu = 0$, contraddicendo l'ipotesi. L'argomento usa la trasformata di Fourier: se l'integrale si annulla per tutti i vettori $\mathbf{w}$, allora la trasformata di Fourier di μ si annulla su un insieme denso, il che per continuità implica $\hat{\mu} \equiv 0$, e dunque $\mu = 0$ per unicità. Il dettaglio tecnico cruciale è che le funzioni $\sigma(\mathbf{w}^T \mathbf{x} + b)$ con $\mathbf{w}$ variabile approssimano arbitrariamente bene le esponenziali complesse $e^{i(\omega^T \mathbf{x})}$, che costituiscono la base della trasformata di Fourier su I .

Osservazioni. Il teorema è esistenziale: garantisce l'esistenza di un'approssimatore ma non fornisce un algoritmo per trovarlo, né bounds sul numero N di unità necessarie. Versioni quantitative (Barron, 1993) mostrano che per funzioni con trasformata di Fourier assolutamente integrabile, l'errore quadratico medio della migliore approssimazione con N unità scala come $O(C_f^2 / N)$, dove $C_f = \int |\omega| |f(\omega)| d\omega$ è una misura della "complessità frequenziale" di f . Hornik (1991) generalizza il teorema a qualsiasi funzione di attivazione non polinomiale, eliminando il requisito di sigmodalità.

Appendice B — La regola della catena in forma generale: differenziale di Fréchet e Jacobiani

La derivazione della backpropagation presentata nel testo usa la regola della catena in forme esplicite per funzioni da $\mathbb{R}^n$ a $\mathbb{R}^m$. Qui formalizziamo il contesto più generale.

Differenziale di Fréchet. Sia $f : U \subseteq \mathbb{R}^n \rightarrow \mathbb{R}^m$ una funzione definita su un aperto U . Si dice che f è *differenziabile in Fréchet* nel punto $\mathbf{x}$ se esiste una trasformazione lineare $D_f(\mathbf{x}) : \mathbb{R}^n \rightarrow \mathbb{R}^m$ tale che:

$$\lim_{\|\mathbf{h}\| \rightarrow 0} \frac{\|f(\mathbf{x} + \mathbf{h}) - f(\mathbf{x}) - D_f(\mathbf{x})\mathbf{h}\|}{\|\mathbf{h}\|} = 0$$

La trasformazione lineare $D_f(\mathbf{x})$ è il *differenziale di Fréchet* di f in $\mathbf{x}$, e la sua matrice rappresentante è il *Jacobiano* $Jf(\mathbf{x}) \in \mathbb{R}^{(m \times n)}$, con (i,j) -esimo elemento $\partial f_i / \partial x_j$.

Regola della catena in forma Jacobiana. Se $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ è differenziabile in $\mathbf{x}$ e $g : \mathbb{R}^m \rightarrow \mathbb{R}^p$ è differenziabile in $\mathbf{y} = f(\mathbf{x})$, allora $h = g \circ f$ è differenziabile in $\mathbf{x}$ e:

$$J(**x**) = Jg(f(**x**)) \cdot Jf(**x**)$$

Come prodotto matriciale: $J \in \mathbb{R}^{(p \times n)} = [J \in \mathbb{R}^{(p \times m)}] \cdot [J \in \mathbb{R}^{(m \times n)}]$.

Applicazione alla backpropagation. Per una rete con composizione $F = f^L \circ f^{L-1} \circ \dots \circ f^1$, dove $f^l(a) = \sigma^l(W^l a + b^l)$, e con perdita scalare $\mathcal{L} = \ell(F(x))$:

$$\begin{aligned} \partial \mathcal{L} / \partial **a^{l-1} &= (Jf^l(**a^{l-1}**))^T \cdot (\partial \mathcal{L} / \partial **a^l) \\ &= (**W^{**l})^T \text{diag}(\sigma'^l(**z^{**l}*)) \cdot (\partial \mathcal{L} / \partial **a^{**l}) \end{aligned}$$

Il Jacobiano di f^l rispetto all'input è:

$$Jf^l(**a^{**l-1}*) = \text{diag}(\sigma'^l(**z^{**l}*)) \cdot **W^{**l} \in \mathbb{R}^{(n^l \times n^{l-1})}$$

Dunque la ricorrenza backward sulle attivazioni è:

$$\partial \mathcal{L} / \partial **a^{**l-1} = (**W^{**l})^T \cdot (\partial \mathcal{L} / \partial **a^{**l} \odot \sigma'^l(**z^{**l}*)) = (**W^{**l})^T **\delta^{**l}$$

che è la forma equivalente (in termini di attivazioni invece di pre-attivazioni) della ricorrenza standard.

Nota sulla non-differenziabilità. La ReLU non è differenziabile in $z = 0$. In pratica, si usano *subgradienti*: un subgradiente g di f nel punto x soddisfa $f(y) \geq f(x) + g \cdot (y - x)$ per tutti y . Per la ReLU, il subgradiente in 0 è qualunque valore in $[0,1]$. La scelta convenzionale $g = 0$ in 0 (oppure $g = 1$) è consistente con la convergenza della SGD nelle condizioni standard (funzione localmente Lipschitz, insieme dei non-differenziabili di misura zero).

Appendice C — Analisi spettrale del vanishing gradient: teorema di Bengio et al.

Bengio, Simard e Frasconi (1994) forniscono un'analisi formale del vanishing/exploding gradient per le RNN, che si generalizza alle reti feedforward profonde. Presentiamo la versione feedforward.

Setup. Consideriamo una rete con L strati, funzioni di attivazione identiche σ , pesi $\mathbf{W}^\ell \in \mathbb{R}^{(n \times n)}$ (reti quadrate per semplicità). Il prodotto dei Jacobiani per gli strati da ℓ a L è:

$$\mathbf{M}_{\ell,L} = \prod_{l=\ell}^L \text{diag}(\sigma'(\mathbf{z}^{**l})) \mathbf{W}^{**l}$$

Il modulo del gradiente allo strato ℓ è $\|\delta^\ell\| \approx \|\mathbf{M}_{\ell+1,L}\|^T \delta^L$.

Lemma del raggio spettrale. Sia $\mathbf{A}$ una matrice quadrata e $\rho(\mathbf{A})$ il suo raggio spettrale (massimo modulo degli autovalori). Per ogni norma matriciale submoltiplicativa $\|\cdot\|$:

$$\|\mathbf{A}^{**k}\|^{1/k} \rightarrow \rho(\mathbf{A}^{**}) \quad \text{per } k \rightarrow \infty$$

Quindi per prodotti di matrici con distribuzione stazionaria:

$$\|\mathbf{M}_{\ell,L}^{**}\|^{1/(L-\ell)} \rightarrow \rho_{\text{effettivo}}$$

dove $\rho_{\text{effettivo}}$ è il raggio spettrale effettivo del prodotto moltiplicativo.

Teorema di Furstenberg-Kesten. Per prodotti di matrici casuali indipendenti $\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_k$ con la stessa distribuzione, vale la legge dei grandi numeri moltiplicativa:

$$(1/k) \log \|\mathbf{A}_1^{**} \mathbf{A}_2^{**} \dots \mathbf{A}_k^{**}\| \rightarrow \gamma_1 \quad \text{q.s. per } k \rightarrow \infty$$

dove γ_1 è il massimo esponente di Lyapunov. Il gradiente $\|\delta^\ell\| \approx e^{\{(L-\ell)\gamma_1\}} \cdot \|\delta^L\|$.

Il gradiente svanisce se $\gamma_1 < 0$ ed esplode se $\gamma_1 > 0$. La condizione di neutralità $\gamma_1 = 0$ è tipicamente non raggiunta spontaneamente e richiede un'inizializzazione accurata (He, Xavier) o meccanismi di normalizzazione esplicita (BatchNorm, LayerNorm).

Implicazione per le LSTM. Le LSTM (Hochreiter e Schmidhuber, 1997) affrontano il problema del vanishing gradient introducendo connessioni di memoria *additive* anziché moltiplicative. Lo stato della cella è:

$$\mathbf{c}^{**t} = \mathbf{f}^{**t} \odot \mathbf{c}^{**_{t-1}} + \mathbf{i}^{**t} \odot \mathbf{g}^{**t}$$

Il gradiente dello stato della cella al tempo s rispetto a quello al tempo $t < s$ è:

$$\partial \mathbf{c}^{**s} / \partial \mathbf{c}^{**t} = \prod_{k=t+1}^s \text{diag}(\mathbf{f}^{**k})$$

che dipende soltanto dai gate di dimenticanza $\mathbf{f}_k \in (0,1)$. Se $\mathbf{f}_k \approx 1$ per tutti i k (gate aperto, nulla viene dimenticato), allora $\partial \mathbf{c}_s / \partial \mathbf{c}_t \approx \mathbf{I}$ — il gradiente non svanisce,

indipendentemente dalla lunghezza della sequenza. Questa è la proprietà chiave che rende le LSTM capaci di apprendere dipendenze a lungo raggio.

Appendice D — Dimostrazione dell'equivalenza predictive coding / backpropagation

La dimostrazione formale di Millidge, Tschantz e Buckley (2020) richiede un framework preciso. Presentiamo la versione per reti feedforward lineari, poi discutiamo l'estensione non lineare.

Modello. Rete feedforward con L strati, funzioni generative lineari $g^l(\mathbf{r}^{l+1}) = \mathbf{W}^l \mathbf{r}^{l+1}$. Le equazioni di inferenza variazionale del predictive coding sono:

$$d\mu^l/dt = -(1/\sigma^{l2}) \epsilon^l + (1/\sigma^{l+12}) (**\mathbf{W}^{**l})^T \epsilon^{l+1}$$

dove $\epsilon^l = \mu^l - \mathbf{W}^l \mu^{l+1}$ è l'errore di previsione. La perdita è $\mathcal{L} = (1/2\sigma^{12}) \|\mu^0 - \mathbf{W}^1 \mu^2\|^2 = (1/2\sigma^{12}) \|\epsilon^0\|^2$ per semplicità.

Condizione di convergenza. Assumiamo $\sigma^l = 1$ per semplicità (precisione omogenea). All'equilibrio ($d\mu^l/dt = 0$ per ogni l):

$$\epsilon^l = (**\mathbf{W}^{**l-1})^T \epsilon^{l-1}$$

Sostituendo ricorsivamente da $l = 1$ verso $l = L$, si ottiene:

$$\epsilon^l = ((**\mathbf{W}^{**l-1})^T ((**\mathbf{W}^{**l-2})^T \dots ((**\mathbf{W}^{**1})^T \epsilon^0) \dots))$$

ossia $\epsilon^l = (\prod_{k=1}^{l-1} (\mathbf{W}^k)^T) \epsilon^0$. Si riconosce che ϵ^l è esattamente il segnale di errore δ^l della backpropagation se $\epsilon^0 = \hat{\mathbf{y}} - \mathbf{y}$ (con attivazioni lineari $\sigma^l=1$).

Aggiornamento dei pesi. Il gradiente della free energy rispetto a $\mathbf{W}^l$ è:

$$\partial F / \partial **\mathbf{W}^{**l} = -\epsilon^l (\mu^{l+1})^T = -**\delta^{**l} (**\mathbf{a}^{**l})^T$$

che è identico al gradiente della backpropagation $\partial \mathcal{L} / \partial \mathbf{W}^l = \delta^l (\mathbf{a}^{l-1})^T$ (con la convenienza di indicizzazione aggiustata). $\square$

Estensione non lineare. Per attivazioni non lineari σ^l , le equazioni di inferenza diventano:

$$d\mu^l/dt = -\sigma'^l(\mu^l) \odot \epsilon^l + Jg^l(**\mu^{**l+1})^T \epsilon^{l+1}$$

dove J_g^l è il Jacobiano della funzione generativa. All'equilibrio, sostituendo e semplificando, si ottiene che il gradiente del predictive coding differisce da quello della backpropagation per termini di ordine $O(\epsilon^2)$ — piccoli quando gli errori di previsione sono piccoli, cioè quando il modello generativo è sufficientemente preciso. Formalmente:

$$\partial F_{PC}/\partial \theta^l = \partial \mathcal{L}_{BP}/\partial \theta^l + O(\|\epsilon\|^2)$$

Appendice E — Derivazione dell'Elastic Weight Consolidation

L'EWC (Kirkpatrick et al., 2017) è derivabile nel framework bayesiano. Dopo aver appreso il compito A con dati $\mathcal{D}_A$, la distribuzione a posteriori sui parametri è $p(\theta \mid \mathcal{D}_A) \propto p(\mathcal{D}_A \mid \theta) p(\theta)$. Per il nuovo compito B con dati $\mathcal{D}_B$, si vuole massimizzare:

$$\log p(\theta \mid \mathcal{D}_A, \mathcal{D}_B) = \log p(\mathcal{D}_B \mid \theta) + \log p(\theta \mid \mathcal{D}_A) + \text{const}$$

Sostituendo $p(\mathcal{D}_B \mid \theta)$ con la likelihood e approssimando $\log p(\theta \mid \mathcal{D}_A)$ con un'espansione di Laplace attorno all'ottimale θ^*_A :

$$\log p(\theta \mid \mathcal{D}_A) \approx \log p(\theta^*_A \mid \mathcal{D}_A) - (1/2)(\theta - \theta^*_A)^T \mathbf{H}^*_{\mathcal{D}_A} (\theta - \theta^*_A)$$

dove $\mathbf{H}_{\mathcal{D}_A} = -\partial^2 \log p(\theta \mid \mathcal{D}_A) / \partial \theta \partial \theta^T \mid \{\theta^*_A\}$ è la matrice di informazione di Hessiana. Approssimando $\mathbf{H}_{\mathcal{D}_A}$ con la matrice di informazione di Fisher diagonale $\mathbf{F}_A$:

$$F_A^{ii} = \mathbb{E}_{\{x^* \sim p(\mathcal{D}_A)\}} [(\partial \log p(y^* \mid x^*, \theta^*_A) / \partial \theta_i)^2]$$

si ottiene la funzione obiettivo EWC:

$$\Delta_{EWC}(\theta) = \Delta_B(\theta) + (\lambda/2) \sum_i F_A^{ii} (\theta_i - \theta^*_{Ai})^2$$

La matrice di Fisher F_A^{ii} approssima la curvatura della perdita A rispetto al parametro θ_i ; parametri con alta curvatura (molto importanti per A) ricevono una penalizzazione più forte se modificati. Il iperparametro λ controlla il compromesso tra fedeltà ai compiti precedenti e adattabilità al compito nuovo.

Connessione con la sinapsi biologica. Il termine F_A^{ii} corrisponde biologicamente alla *sensibilità* della sinapsi all'aggiornamento — la quantità di informazione che quella sinapsi porta sulla distribuzione di input del compito A. Sinapsi ad alta informazione corrispondono a connessioni consolidate; sinapsi a bassa informazione corrispondono a connessioni plastiche. Il meccanismo di PKM ζ che protegge le sinapsi consolidate dal decadimento implementa biologicamente una penalizzazione che scala con l'importanza della sinapsi — non esattamente la Fisher information, ma una quantità funzionalmente analoga derivata dalla storia di attività.

Appendice F — Teoria dell'informazione e capacità di apprendimento

Entropia di Shannon e informazione mutua. Per variabili aleatorie discrete X e Y , l'entropia e l'informazione mutua sono:

$$H(X) = -\sum_x p(x) \log p(x)$$

$$H(X|Y) = -\sum_{\{x,y\}} p(x,y) \log p(x|y)$$

$$I(X;Y) = H(X) - H(X|Y) = \sum_{\{x,y\}} p(x,y) \log [p(x,y)/(p(x)p(y))]$$

Information Bottleneck e apprendimento delle rappresentazioni. Tishby e Schwartz-Ziv (2017) propongono di analizzare l'apprendimento nelle reti neurali attraverso il piano informativo ($I(X;T)$, $I(T;Y)$), dove T è la rappresentazione interna di uno strato, X è l'input e Y è il target. Il principio dell'Information Bottleneck (IB) definisce la rappresentazione ottimale come quella che massimizza $I(T;Y)$ (informazione sul target) minimizzando $I(X;T)$ (complessità della rappresentazione):

$$\min_{\{p(t|x)\}} I(X;T) - \beta I(T;Y)$$

per un parametro di trade-off $\beta > 0$. Tishby sostiene che il training con backpropagation attraversa due fasi nel piano IB: nella prima fase (fitting), $I(X;T)$ e $I(T;Y)$ crescono entrambi mentre la rete "memorizza" i dati; nella seconda fase (compression), $I(X;T)$ si riduce mentre $I(T;Y)$ rimane stabile, indicando che la rete "dimentica" le informazioni irrilevanti sull'input.

Queste osservazioni, sebbene contestate da Saxe et al. (2019) che mostrano la dipendenza dalla scelta del binning per la stima dell'informazione, sono intellettualmente rilevanti per il confronto biologico: il cervello è noto per codificare in modo sparso e compresso, consistentemente con un obiettivo di minimizzazione della complessità rappresentazionale. Il principio IB potrebbe essere una formalizzazione matematica del principio di codifica efficiente di Barlow (1961), applicato all'apprendimento supervisionato.

Capacità di memoria delle reti associative. Per una rete di Hopfield con N neuroni binari (± 1), la capacità di memoria massima — il numero di pattern p che possono essere memorizzati e recuperati senza errori — è:

$$p_{\max} \approx 0.138 N \quad (\text{Cover, 1988; Amit et al., 1985})$$

Con codifica sparsa (attivazione di solo $k \ll N$ neuroni per pattern), la capacità scala come:

$$p_{\max} \approx N / (k \log(N/k))$$

che per $k = \alpha N$ con $\alpha \rightarrow 0$ cresce come $N/(\alpha N \log(1/\alpha)) = 1/(\alpha \log(1/\alpha))$, divergendo quando $\alpha \rightarrow 0$. La sparsità aumenta la capacità in modo superlineare — il fondamento

teorico per cui l'ippocampo, con le sue rappresentazioni sparsissime nel giro dentato ($< 5\%$ di neuroni attivi per pattern), può memorizzare un numero enormemente superiore di episodi rispetto a strutture con codifica densa.

Appendice G — Il principio del massimo di Pontryagin e la connessione con la backpropagation

La backpropagation è un caso speciale del metodo degli stati aggiunti (*adjoint state method*) usato nell'ottimizzazione dei sistemi dinamici continui. Questa connessione, riconosciuta da Werbos (1974) nella sua dissertazione, è raramente esposta nei testi di deep learning moderni.

Problema di controllo ottimo. Consideriamo un sistema dinamico:

$$d\mathbf{x}^{**}(t)/dt = \mathbf{f}(\mathbf{x}^{**}(t), \mathbf{u}^{**}(t), t) \quad \mathbf{x}^{**}(0) = \mathbf{x}^{**}_0$$

con variabile di stato $\mathbf{x}(t) \in \mathbb{R}^n$ e controllo $\mathbf{u}(t) \in \mathbb{R}^m$. Si vuole minimizzare il costo:

$$J = \Phi(\mathbf{x}^{**}(T)) + \int_0^T L(\mathbf{x}^{**}(t), \mathbf{u}^{**}(t), t) dt$$

Il principio del massimo di Pontryagin introduce le variabili aggiunte (costati) $\boldsymbol{\lambda}(t) \in \mathbb{R}^n$ che soddisfano:

$$d\boldsymbol{\lambda}^{**}(t)/dt = -\partial H / \partial \mathbf{x}^{**} \quad \boldsymbol{\lambda}^{**}(T) = \partial \Phi / \partial \mathbf{x}^{**}(T)$$

dove $H(\mathbf{x}, \mathbf{u}, \boldsymbol{\lambda}, t) = \boldsymbol{\lambda}^T \mathbf{f}(\mathbf{x}, \mathbf{u}, t) - L(\mathbf{x}, \mathbf{u}, t)$ è l'hamiltoniana del sistema. La condizione di ottimalità rispetto al controllo è:

$$\partial H / \partial \mathbf{u}^{**} = 0 \Rightarrow \boldsymbol{\lambda}^{**T} \partial \mathbf{f} / \partial \mathbf{u}^{**} = \partial L / \partial \mathbf{u}^{**}$$

Analogia con la backpropagation. La rete neurale feedforward è un sistema dinamico discreto con "tempo" $\ell = 1, \dots, L$:

$$\mathbf{a}^{**\ell} = \mathbf{f}(\mathbf{a}^{**\ell-1}, \mathbf{W}^{**\ell}) = \sigma(\mathbf{W}^{**\ell} \mathbf{a}^{**\ell-1} + \mathbf{b}^{**\ell})$$

I "controlli" sono i parametri $\mathbf{W}^\ell, \mathbf{b}^\ell$; il "costo terminale" è $\mathcal{L}(\mathbf{a}^L, \mathbf{y})$. Le variabili aggiunte sono $\boldsymbol{\lambda}^\ell = \partial \mathcal{L} / \partial \mathbf{a}^\ell$, e le equazioni degli stati aggiunti sono:

$$\boldsymbol{\lambda}^{**\ell-1} = (\partial \mathbf{f} / \partial \mathbf{a}^{**\ell-1})^T \boldsymbol{\lambda}^{**\ell} = (\mathbf{J} \mathbf{f})^T \boldsymbol{\lambda}^{**\ell} \quad \text{con} \quad \boldsymbol{\lambda}^{**L} = \partial \mathcal{L} / \partial \mathbf{a}^{**L}$$

che è esattamente la ricorrenza backward della backpropagation. La condizione di ottimalità rispetto ai parametri $\mathbf{W}^l$ è:

$$\partial \mathcal{L} / \partial \mathbf{W}^{**l} = (\partial f / \partial \mathbf{W}^{**l})^T \mathbf{\lambda}^{**l} = \mathbf{\lambda}^{**l} (\mathbf{a}^{**l-1})^T = \mathbf{\delta}^{**l} (\mathbf{a}^{**l-1})^T$$

La backpropagation è dunque la discretizzazione delle equazioni degli stati aggiunti di Pontryagin per un sistema di controllo ottimo a tempo discreto. Questa prospettiva chiarisce perché la backpropagation calcola i gradienti *esatti* (non approssimati): le equazioni degli stati aggiunti sono la condizione necessaria di primo ordine per l'ottimalità, e il metodo dei moltiplicatori di Lagrange garantisce che la soluzione di queste equazioni produca il gradiente esatto della funzione di costo rispetto ai parametri.

Appendice H — Free Energy Principle: derivazione variazionale completa

Il *free energy principle* di Friston è formalizzato come la minimizzazione della free energy variazionale F rispetto a una distribuzione di riconoscimento $q(\mathbf{z} | \mathbf{x})$ che approssima la vera posteriori $p(\mathbf{z} | \mathbf{x})$.

Definizione. Per variabili latenti $\mathbf{z}$ e osservazioni $\mathbf{x}$:

$$\begin{aligned} F[q] &= \int q(\mathbf{z}) \log [q(\mathbf{z}) / p(\mathbf{x}, \mathbf{z})] d\mathbf{z} \\ &= \text{KL}[q(\mathbf{z}) \parallel p(\mathbf{z} | \mathbf{x})] - \log p(\mathbf{x}) \\ &\geq -\log p(\mathbf{x}) \quad (\text{per non-negatività della KL}) \end{aligned}$$

Minimizzare F rispetto a q soddisfa due obiettivi: riduce la KL dalla posteriori vera (inferenza accurata) e fornisce un lower bound variazionale alla log-verosimiglianza (equivalente all'ELBO — Evidence Lower BOund — nella letteratura delle variational autoencoders).

Decomposizione alternativa. La free energy si decompone anche come:

$$\begin{aligned} F[q] &= \mathbb{E}_q[-\log p(\mathbf{x} | \mathbf{z})] + \text{KL}[q(\mathbf{z}) \parallel p(\mathbf{z})] \\ &= \text{complessità} + \text{accuratezza} \end{aligned}$$

Il termine di complessità $\text{KL}[q \parallel p(\mathbf{z})]$ misura quanto q si discosta dal prior; il termine di accuratezza misura l'errore di ricostruzione. Minimizzare F bilancia l'adattamento ai dati (riduzione dell'accuratezza) con la semplicità della rappresentazione (riduzione della complessità), implementando computazionalmente il principio del rasoio di Occam.

Inferenza attiva. Friston estende il FEP all'azione: l'agente può modificare non solo le proprie credenze ($\mathbf{z}$) ma anche lo stato del mondo ($\mathbf{x}$) mediante azioni. Azioni che riducono la surprise attesa — portando le osservazioni in accordo con le previsioni — minimizzano F sull'intera storia futura. Formalmente:

$$**a** = \arg \min_{**a**} \mathbb{E}_{\{p(**x**_{t+1}) | **a**, **x**_t\}} [F(q(**z**_{t+1}) | **x**_{t+1})]$$

Questa formulazione unifica la percezione (aggiornamento di q) e l'azione (scelta di a) come aspetti del medesimo processo di minimizzazione della free energy, con implicazioni per la teoria del comportamento intenzionale e della cognizione embodied.

Connessione con la backpropagation. Come mostrato nell'Appendice D, il predictive coding — il meccanismo computazionale attraverso cui il FEP è implementato nella gerarchia corticale — converge alla backpropagation nel limite di inferenza perfetta. Più in generale, qualunque sistema che minimizzi la free energy variazionale F rispetto ai parametri del modello generativo $p(\mathbf{x}, \mathbf{z}; \theta)$ esegue implicitamente la discesa del gradiente su un funzionale della log-verosimiglianza marginale, producendo aggiornamenti dei parametri che approssimano quelli della backpropagation nel caso supervisionato.

Appendice I — Grounding della STDP nella Spike Response Model

La STDP come regola fenomenologica emerge naturalmente dalle proprietà biofisiche dei potenziali d'azione in modelli di neuroni spiking. La *Spike Response Model* (SRM) di Gerstner e Kistler (2002) descrive il potenziale di membrana $u(t)$ come la sovrapposizione di contributi filtrati:

$$u(t) = \sum_j \sum_i w_{ij} \epsilon_{ij}(t - t^s_j) + \sum_i \eta(t - t^s_i) + u_{\text{rest}}$$

dove $\epsilon_{ij}(t - t^s_j)$ è il profilo temporale del potenziale post-sinaptico eccitatorio (EPSP) indotto dallo spike del neurone j al tempo t^s_j , $\eta(t - t^s_i)$ è il potenziale di afterhyperpolarization successivo a uno spike del neurone i al tempo t^s_i , e w_{ij} è il peso sinaptico.

Nel framework SRM, la probabilità di firing del neurone i nell'intervallo $[t, t+dt)$ è:

$$P(\text{spike in } [t, t+dt)) = \rho(u(t) - \vartheta) dt$$

dove ρ è una funzione di intensità sigmoideale e ϑ è la soglia. Massimizzando la log-verosimiglianza della sequenza osservata di spike rispetto ai pesi w_{ij} mediante gradient ascent:

$$\Delta w_{ij} \propto \partial \log P(\text{spike train}) / \partial w_{ij}$$

si ottiene, dopo calcoli, la regola:

$$\Delta w_{ij} = \eta \sum_s \{ \text{spike } t^s_j \} [\tilde{\epsilon}(t^s_j - \text{ultima volta che } j \text{ ha spikato}) - \text{baseline}]$$

dove $\tilde{\epsilon}$ è un filtro temporale. Questa espressione, nella forma del suo profilo temporale, riproduce qualitativamente la finestra STDP: se j spika prima di i , la sinapsi si rafforza (LTP); se j spika dopo i , la sinapsi si indebolisce (LTD). La STDP è dunque la regola di gradient ascent sulla log-verosimiglianza dei spike nell'ambito del modello SRM — un risultato che la ancora saldamente al framework dell'inferenza probabilistica senza richiedere la propagazione esplicita di gradienti globali.

Appendice L — Misure di similarità tra rappresentazioni: RSA e CKA

Nella Parte III è stato discusso l'uso della RSA (Representational Similarity Analysis) per confrontare attività neuronali e attivazioni artificiali. Qui formalizziamo la RSA e introduciamo la metrica alternativa CKA.

RSA. Per un set di N stimoli, la rappresentazione del sistema A produce una matrice $\text{RDM}_A \in \mathbb{R}^{(N \times N)}$:

$$\text{RDM}_A[i, j] = 1 - \rho(\mathbf{r}_A(i), \mathbf{r}_A(j))$$

dove $\mathbf{r}_A(i)$ è il vettore di risposta al stimolo i e ρ è la correlazione di Pearson. La similarità tra i sistemi A e B è:

$$\rho_{\text{RSA}}(A, B) = \rho_{\text{Spearman}}(\text{vec}(\text{RDM}_A), \text{vec}(\text{RDM}_B))$$

dove $\text{vec}()$ estrae la parte triangolare superiore. La correlazione di Spearman è usata per robustezza rispetto agli outlier nelle dissimilarità.

CKA (Centered Kernel Alignment). Kornblith et al. (2019) propongono una metrica alternativa che supera alcune limitazioni della RSA, in particolare la dipendenza dal numero di stimoli e la sensibilità al rumore nelle singole rappresentazioni. Per due matrici di attivazioni $\mathbf{X} \in \mathbb{R}^{(N \times p)}$ e $\mathbf{Y} \in \mathbb{R}^{(N \times q)}$, le matrici di kernel centrate sono:

$$\mathbf{K}_X = \mathbf{H} \mathbf{X} \mathbf{X}^T \mathbf{H} \quad \mathbf{K}_Y = \mathbf{H} \mathbf{Y} \mathbf{Y}^T \mathbf{H}$$

dove $\mathbf{H} = \mathbf{I} - (1/N) \mathbf{1} \mathbf{1}^T$ è la matrice di centratura. La CKA è:

$$\begin{aligned} \text{CKA}(\mathbf{X}, \mathbf{Y}) &= \|\mathbf{K}_X - \mathbf{K}_Y\|_F / (\|\mathbf{K}_X\|_F \|\mathbf{K}_Y\|_F) \\ &= \text{HSIC}(\mathbf{X}, \mathbf{Y}) / \sqrt{(\text{HSIC}(\mathbf{X}, \mathbf{X}) \cdot \text{HSIC}(\mathbf{Y}, \mathbf{Y}))} \end{aligned}$$

dove HSIC è la Hilbert-Schmidt Independence Criterion. La CKA è invariante a trasformazioni ortogonali e a riscaldamento isotropo delle rappresentazioni, il che la rende più appropriata per confrontare rappresentazioni di dimensionalità diversa (come neuroni biologici e unità artificiali).

Kornblith et al. mostrano che la CKA è più coerente della RSA nel rilevare la corrispondenza tra strati di reti addestrate con architettura identica ma parametri diversi, e più robusta al campionamento casuale dei neuroni. Tuttavia, sia RSA sia CKA sono misure di seconda distanza — confrontano strutture di (dis)similarità piuttosto che vettori di risposta — e quindi non catturano informazione assoluta sulla direzione delle rappresentazioni, il che è rilevante quando si vuole verificare se i segnali di errore si propagano nella direzione corretta.

Bibliografia

La bibliografia è organizzata per aree tematiche corrispondenti ai capitoli principali della monografia, con annotazioni sintetiche che indicano il contributo specifico di ciascun lavoro alla discussione. I testi sono citati nella forma autore-anno secondo la convenzione della letteratura scientifica internazionale.

I. Fondamenti storici e filosofici

Minsky M. e Papert S. (1969). *Perceptrons: An Introduction to Computational Geometry*. MIT Press, Cambridge MA. — Analisi rigorosa dei limiti dei perceptron a strato singolo; formula per la prima volta il problema dell'attribuzione del credito negli strati nascosti come ostacolo fondamentale.

Werbos P. (1974). *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences*. Tesi di dottorato, Harvard University. — Prima derivazione dell'algoritmo di backpropagation nel contesto della teoria del controllo ottimo; contiene la connessione formale con il metodo degli stati aggiunti di Pontryagin.

Rumelhart D. E., Hinton G. E. e Williams R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323, 533–536. — La formulazione moderna della backpropagation che ha reso l'algoritmo accessibile alla comunità di machine learning; introduce il concetto di "rappresentazioni distribuite" negli strati nascosti.

Parker D. B. (1985). *Learning Logic*. Technical Report TR-47, Center for Computational Research in Economics and Management Science, MIT. — Derivazione indipendente della backpropagation, contemporanea al lavoro di Rumelhart et al.

LeCun Y. (1986). *Learning processes in an asymmetric threshold network*. In E. Bienenstock, F. Fogelman-Soulié e G. Weisbuch (a cura di), *Disordered Systems and Biological Organization*, pp. 233–240. Springer. — Prima formulazione della target propagation come alternativa biologicamente motivata alla backpropagation.

Marr D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W.H. Freeman, San Francisco. — Propone la tripartizione computazionale/algoritmico/implementazionale che rimane il framework più utile per analizzare il rapporto tra algoritmi di apprendimento e substrati fisici.

Bellman R. (1957). *Dynamic Programming*. Princeton University Press. — Introduce la programmazione dinamica, il framework matematico di cui la backpropagation è un caso speciale.

Pontryagin L. S., Boltyanskii V. G., Gamkrelidze R. V. e Mishchenko E. F. (1962). *The Mathematical Theory of Optimal Processes*. Interscience Publishers, New York. — Il principio del massimo che formalizza le equazioni degli stati aggiunti, di cui la backpropagation è la versione discreta.

II. Algoritmi di apprendimento: teoria matematica

Cybenko G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, 2(4), 303–314. — Teorema di approssimazione universale per reti a uno strato nascosto con attivazioni sigmoidi.

Hornik K., Stinchcombe M. e White H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5), 359–366. — Generalizzazione del teorema di approssimazione universale a qualunque funzione di attivazione non polinomiale.

Barron A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3), 930–945. — Bounds quantitativi sulla capacità di approssimazione in funzione del numero di unità; connessione con la complessità frequenziale dell'input.

Glorot X. e Bengio Y. (2010). Understanding the difficulty of training deep feedforward neural networks. *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 249–256. — Analisi sistematica del vanishing gradient; derivazione dell'inizializzazione Xavier che bilancia la varianza delle attivazioni e dei gradienti.

He K., Zhang X., Ren S. e Sun J. (2015). Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 1026–1034. — Inizializzazione ottimale per ReLU; introduce le ResNet con connessioni residue che risolvono il vanishing gradient in reti molto profonde.

Bengio Y., Simard P. e Frasconi P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2), 157–166. — Analisi formale del vanishing/exploding gradient nelle RNN; formalizzazione degli esponenti di Lyapunov come misura del comportamento asintotico del gradiente.

Hochreiter S. e Schmidhuber J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. — Introduzione delle LSTM come soluzione al vanishing gradient nelle RNN; la

connessione di cella additiva evita il decadimento esponenziale del gradiente attraverso il tempo.

Ioffe S. e Szegedy C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 448–456. — Normalizzazione delle pre-attivazioni per strato; mantiene il raggio spettrale effettivo vicino a 1 durante l'addestramento.

Kingma D. P. e Ba J. (2015). Adam: A method for stochastic optimization. *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*. — Ottimizzatore adattivo che combina momento di primo e secondo ordine; connessione con la natural gradient descent di Amari.

Amari S. (1998). Natural gradient works efficiently in learning. *Neural Computation*, 10(2), 251–276. — Derivazione della discesa del gradiente naturale che corregge per la geometria riemanniana dello spazio dei parametri; fondamento teorico degli ottimizzatori adattativi.

Pascanu R., Mikolov T. e Bengio Y. (2013). On the difficulty of training recurrent neural networks. *Proceedings of the 30th International Conference on Machine Learning (ICML)*, 1310–1318. — Analisi formale del gradiente esplosivo; propone il gradient clipping come soluzione pratica.

Hendrycks D. e Gimpel K. (2016). Gaussian error linear units (GELUs). *arXiv preprint*, arXiv:1606.08415. — Introduzione della GELU; interpretazione come aspettativa stocastica dell'applicazione di ReLU con rumore gaussiano.

Ramachandran P., Zoph B. e Le Q. V. (2017). Searching for activation functions. *arXiv preprint*, arXiv:1710.05941. — Ottimizzazione automatica delle funzioni di attivazione; la funzione Swish emerge come la migliore su diversi benchmark.

III. Plausibilità biologica: teoria e proposte alternative

Crick F. (1989). The recent excitement about neural networks. *Nature*, 337, 129–132. — Prima formulazione sistematica delle obiezioni alla plausibilità biologica della backpropagation; introduce il problema del trasporto dei pesi.

Lillicrap T. P., Cownden D., Tweed D. B. e Akerman C. J. (2016). Random synaptic feedback weights support error backpropagation for deep learning. *Nature Communications*, 7, 13276. — Introduce il feedback alignment; dimostra che matrici di feedback casuali fisse permettono l'apprendimento in reti non banali.

Nøkland A. (2016). Direct feedback alignment provides learning in deep neural networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 29. — Estende il feedback alignment al direct feedback alignment in cui ogni strato riceve il segnale di errore direttamente dall'uscita.

Lee D. H., Zhang S., Fischer A. e Bengio Y. (2015). Difference target propagation. *Proceedings of the European Conference on Machine Learning (ECML-PKDD)*, 498–515. — Introduce la difference target propagation come algoritmo di apprendimento locale; dimostra la connessione con un'approssimazione del passo di Newton.

Bengio Y. (2014). How auto-encoders could provide credit assignment in deep networks via target propagation. *arXiv preprint*, arXiv:1407.7906. — Formalizzazione moderna della target propagation; connessione con l'inferenza variazionale.

Rao R. P. N. e Ballard D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1), 79–87. — Formulazione influente del predictive coding come modello computazionale della gerarchia visiva; introduce gli errori di previsione come segnali da trasmettere verso l'alto.

Millidge B., Tschantz A. e Buckley C. L. (2020). Predictive coding approximates backpropagation along arbitrary computation graphs. *arXiv preprint*, arXiv:2006.04198. — Dimostrazione formale dell'equivalenza tra predictive coding e backpropagation nel limite di convergenza dell'inferenza variazionale.

Sacramento J., Costa R. P., Bengio Y. e Senn W. (2018). Dendritic cortical microcircuits approximate the backpropagation algorithm. *Advances in Neural Information Processing Systems (NeurIPS)*, 31. — Il modello burstprop: computazione bicompartimentale come approssimazione locale della backpropagation; connessione con la fisiologia dendritica dei neuroni piramidali.

Scellier B. e Bengio Y. (2017). Equilibrium propagation: Bridging the gap between energy-based models and backpropagation. *Frontiers in Computational Neuroscience*, 11, 24. — Introduce l'equilibrium propagation; dimostra il teorema centrale sulla convergenza al gradiente esatto per $\beta \rightarrow 0$.

Whittington J. C. R. e Bogacz R. (2017). An approximation of the error backpropagation algorithm in a predictive coding network with local Hebbian synaptic plasticity. *Neural Computation*, 29(5), 1229–1262. — Dimostrazione che il predictive coding con un'unica copia dei pesi (no simmetria) approssima la backpropagation per reti ricorrenti.

Jaderberg M., Czarnecki W. M., Osindero S., Vinyals O., Graves A., Silver D. e Kavukcuoglu K. (2017). Decoupled neural interfaces using synthetic gradients. *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 1627–1635. — Propone gradienti sintetici per risolvere il problema del blocco modulante; ogni strato predice il proprio gradiente senza attendere il backward pass completo.

Miconi T., Clune J. e Stanley K. O. (2018). Differentiable plasticity: Training plastic neural networks with backpropagation. *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 3559–3568. — Plasticità hebbiana apprendibile per gradient descent; l'adattamento rapido durante il compito non richiede backpropagation.

Friston K. J. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2), 127–138. — Formulazione del free energy principle come framework

unificante per percezione, azione e apprendimento; generalizzazione del predictive coding a tutti i livelli di organizzazione neuronale.

IV. Plasticità sinaptica biologica

Hebb D. O. (1949). *The Organization of Behavior: A Neuropsychological Theory*. Wiley, New York. — Formulazione della regola hebbiana; propone che la correlazione tra attività pre- e post-sinaptica sia il meccanismo di base dell'apprendimento associativo.

Bliss T. V. P. e Lomo T. (1973). Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path. *Journal of Physiology*, 232(2), 331–356. — Scoperta sperimentale della Long-Term Potentiation nell'ippocampo; primo documento della plasticità sinaptica persistente indotta dall'attività.

Bi G. e Poo M. (1998). Synaptic modifications in cultured hippocampal neurons: Dependence on spike timing, synaptic strength, and postsynaptic cell type. *Journal of Neuroscience*, 18(24), 10464–10472. — Prima documentazione quantitativa della STDP in neuroni ippocampali in coltura; misura precisa delle costanti di tempo della finestra di plasticità.

Markram H., Lübke J., Frotscher M. e Sakmann B. (1997). Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. *Science*, 275(5297), 213–215. — Dimostrazione della STDP in neuroni piramidali della corteccia somatosensoriale di ratto; coppia pre → post produce LTP, post → pre produce LTD.

Bienenstock E. L., Cooper L. N. e Munro P. W. (1982). Theory for the development of neuron selectivity: Orientation specificity and binocular interaction in visual cortex. *Journal of Neuroscience*, 2(1), 32–48. — Introduzione della regola BCM con soglia scorrevole; spiega la selettività di orientamento in V1 e predice l'omeostasi sinaptica.

Oja E. (1982). A simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, 15(3), 267–273. — Regola hebbiana stabilizzata che converge alla prima componente principale; fondamento della codifica efficiente non supervisionata.

Turrigiano G. G. (2008). The self-tuning neuron: Synaptic scaling of excitatory synapses. *Cell*, 135(3), 422–435. — Revisione sistematica della *synaptic scaling* come meccanismo di stabilizzazione omeostatica; tutte le sinapsi di un neurone vengono scalate moltiplicativamente per mantenere il firing rate target.

Bittner K. C., Milstein A. D., Grienberger C., Romani S. e Magee J. C. (2017). Behavioral time scale synaptic plasticity underlies CA1 place fields. *Science*, 357(6355), 1033–1036. — Scoperta della BTSP nell'ippocampo; i plateau potentials producono LTP massiccia di sinapsi attive fino a 400 ms prima — meccanismo di apprendimento one-shot con attribuzione del credito temporale retrocausale.

Pfister J. P. e Gerstner W. (2006). Triplets of spikes in a model of spike timing-dependent plasticity. *Journal of Neuroscience*, 26(38), 9673–9682. — Regola STDP a tre fattori (triple di spike) che supera le limitazioni della regola a coppie; connessione con la backpropagation through time.

Sacktor T. C. (2011). How does PKM ζ maintain long-term memory? *Nature Reviews Neuroscience*, 12(1), 9–15. — Revisione del ruolo di PKM ζ nella consolidazione sinaptica; meccanismo biologico di protezione delle sinapsi consolidate analogo all'EWC.

V. Neuroscienza computazionale: segnali di errore e apprendimento per rinforzo

Schultz W., Dayan P. e Montague P. R. (1997). A neural substrate of prediction and reward. *Science*, 275(5306), 1593–1599. — Dimostrazione quantitativa che i neuroni dopaminergici della SNc e VTA codificano il TD error (reward prediction error); base sperimentale per la neuroscienze computazionale del rinforzo.

Rescorla R. A. e Wagner A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black e W. F. Prokasy (a cura di), *Classical Conditioning II: Current Research and Theory*, pp. 64–99. Appleton-Century-Crofts. — Il modello Rescorla-Wagner come precursore del TD-learning; la prima formulazione matematica dell'errore di previsione nella condizionamento classico.

Sutton R. S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning*, 3(1), 9–44. — Formulazione del TD-learning e dimostrazione della sua convergenza; connessione con la programmazione dinamica di Bellman.

Sutton R. S. e Barto A. G. (2018). *Reinforcement Learning: An Introduction* (2a ed.). MIT Press. — Il testo di riferimento per l'apprendimento per rinforzo; include trattazione delle three-factor rules biologiche.

Montague P. R., Dayan P. e Sejnowski T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience*, 16(5), 1936–1947. — Connessione formale tra la risposta dopaminergica e il TD-learning; base per l'interpretazione del sistema dopaminergico come implementazione biologica del critic nel framework actor-critic.

O'Doherty J., Dayan P., Schultz J., Deichmann R., Friston K. e Dolan R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science*, 304(5669), 452–455. — Dissociazione sperimentale mediante fMRI tra striato ventrale (critic, aggiornamento del valore) e striato dorsale (actor, selezione dell'azione) nel framework actor-critic.

VI. Evidenza sperimentale: visione, gerarchia corticale e oscillazioni

Hubel D. H. e Wiesel T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, 160(1), 106–154. — Scoperta delle cellule semplici e complesse in V1; fondamento dell'organizzazione gerarchica della corteccia visiva.

Yamins D. L. K., Hong H., Cadieu C. F., Solomon E. A., Seibert D. e DiCarlo J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23), 8619–8624. — Prima dimostrazione sistematica che le CNN ottimizzate per compiti visivi predicono le risposte di IT della scimmia con accuratezza superiore a qualunque modello precedente.

Kriegeskorte N., Mur M. e Bandettini P. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2, 4. — Introduzione della RSA come metrica di confronto tra rappresentazioni di sistemi eterogenei; metodo diventato standard nel campo.

Bastos A. M., Usrey W. M., Adams R. A., Mangun G. R., Fries P. e Friston K. J. (2012). Canonical microcircuits for predictive coding. *Neuron*, 76(4), 695–711. — Dissociazione spettrale feedforward/feedback (gamma vs. beta) mediante Granger causality; proposta del circuito canonico di microstrato per il predictive coding.

Schoups A., Vogels R., Qian N. e Orban G. (2001). Practising orientation identification improves orientation coding in V1 neurons. *Nature*, 412(6845), 549–553. — Modifiche specifiche di orientamento in V1 dopo apprendimento discriminativo; evidenza di apprendimento guidato da segnali top-down in strati profondi della gerarchia.

Murray S. O., Kersten D., Olshausen B. A., Schrater P. e Woods D. L. (2002). Shape perception reduces activity in human primary visual cortex. *Proceedings of the National Academy of Sciences*, 99(23), 15164–15169. — Soppressione delle risposte corticali per stimoli predetti; evidenza diretta di segnali predittivi nella corteccia visiva.

Wilson M. A. e McNaughton B. L. (1994). Reactivation of hippocampal ensemble memories during sleep. *Science*, 265(5172), 676–679. — Prima documentazione del replay sequenziale delle cellule di posto durante il sonno; fondamento sperimentale per la teoria del consolidamento ippocampo-neocorteccia.

Diba K. e Buzsáki G. (2007). Forward and reverse hippocampal place-cell sequences during ripples. *Nature Neuroscience*, 10(10), 1241–1242. — Distinzione tra forward replay (ordine dell'esplorazione) e reverse replay (ordine inverso) durante i sharp-wave ripples; interpretazione come propagazione backward del valore.

VII. Apprendimento nei modelli biologicamente vincolati

McClelland J. L., McNaughton B. L. e O'Reilly R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102(3), 419–457. — La

teoria dei sistemi di apprendimento complementari; la dicotomia ippocampo/neocorteccia come soluzione biologica al catastrophic forgetting.

Kirkpatrick J., Pascanu R., Rabinowitz N., Veness J., Desjardins G., Rusu A. A., Milan K., Quan J., Ramalho T., Grabska-Barwinska A., Hassabis D., Clopath C., Kumaran D. e Hadsell R. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13), 3521–3526. — Introduzione dell'EWC; derivazione bayesiana della penalizzazione dei parametri importanti per compiti precedenti; connessione con la consolidazione sinaptica biologica.

Finn C., Abbeel P. e Levine S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 1126–1135. — Il framework MAML per il meta-apprendimento; la ricerca dell'inizializzazione ottimale come formalizzazione computazionale dell'adattamento evolutivo.

Mnih V., Kavukcuoglu K., Silver D., Rusu A. A., Veness J., Bellemare M. G., Graves A., Riedmiller M., Fidjeland A. K., Ostrovski G., Petersen S., Beattie C., Sadik A., Antonoglou I., King H., Kumaran D., Wierstra D., Legg S. e Hassabis D. (2015). Human-level control through deep reinforcement learning. *Nature*, 518, 529–533. — Deep Q-Network con experience replay; ispirazione dal replay ippocampale; raggiungimento di performance umana su 49 giochi Atari.

Zador A. M. (2019). A critique of pure learning and what artificial neural networks can learn from animal brains. *Nature Communications*, 10, 3770. — Argomento per il ruolo primario delle strutture innate (connettoma geneticamente determinato) rispetto all'apprendimento esperienziale; critica del paradigma "tabula rasa" nel deep learning.

Marr D. e Albus J. S. (1971). A theory of cerebellar cortex. *Journal of Physiology*, 202(2), 437–470. (Marr 1969; Albus 1971). — Modello computazionale dell'apprendimento cerebellare; le fibre rampicanti come segnale di errore per la modifica LTD delle sinapsi PF → PC.

Ito M., Sakurai M. e Tongroach P. (1982). Climbing fibre induced depression of both mossy fibre responsiveness and glutamate sensitivity of cerebellar Purkinje cells. *Journal of Physiology*, 324(1), 113–134. — Dimostrazione sperimentale della LTD cerebellare indotta dalla co-attivazione di fibre parallele e rampicanti; fondamento sperimentale del modello Marr-Albus.

VIII. Neuromorphic computing e hardware biologicamente ispirato

Davies M., Srinivasa N., Lin T. H., Chinya G., Cao Y., Choday S. H., Dimou G., Joshi P., Imam N., Jain S., Liao Y., Lin C. K., Lines A., Liu R., Mathaikutty D., McCoy S., Paul A., Tse J., Venkataramanan G., Weng Y. T., Wild A., Yang Y. e Wang H. (2018). Loihi: A neuromorphic manycore processor with on-chip learning. *IEEE Micro*, 38(1), 82–99. —

Descrizione del chip Intel Loihi; implementazione on-chip di neuroni spiking con plasticità programmabile incluse varianti STDP.

Merolla P. A., Arthur J. V., Alvarez-Icaza R., Cassidy A. S., Sawada J., Akopyan F., Jackson B. L., Imam N., Guo C., Nakamura Y., Brezzo B., Vo I., Esser S. K., Appuswamy R., Taba B., Amir A., Flickner M. D., Risk W. P., Manohar R. e Modha D. S. (2014). A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science*, 345(6197), 668–673. — Il chip IBM TrueNorth con 1 milione di neuroni sintetici; architettura event-driven a basso consumo energetico.

Scellier B., Goyal A., Binas J., Mesnard T., Pezzulo G. e Bengio Y. (2023). Agnostic local learning rules for efficient training of deep networks with equilibrium propagation. *arXiv preprint*, arXiv:2306.02644. — Implementazione dell'equilibrium propagation su circuiti analogici; dimostrazione di prestazioni competitive su CIFAR-10 con regole di apprendimento locali.

IX. Intelligenza artificiale e neuroscienze: frontiere aperte

Schrimpf M., Blank I., Tuckute G., Kauf C., Hosseini E. A., Kanwisher N., Tenenbaum J. e Fedorenko E. (2020). The neural architecture of language: Integrative reverse-engineering perspectives from machine learning. *bioRxiv*, 2020.06.26.174482. — Confronto sistematico tra transformer e attività corticale durante la lettura; gli strati intermedi di BERT predicono le risposte di aree linguistiche meglio dei modelli precedenti.

Tishby N. e Schwartz-Ziv R. (2017). Opening the black box of deep neural networks via information. *arXiv preprint*, arXiv:1703.00810. — Il piano informativo per analizzare l'apprendimento nelle reti neurali; le fasi di fitting e compression nel piano $I(X;T)-I(T;Y)$.

Dehaene S., Changeux J. P. e Naccache L. (2011). The global neuronal workspace model of conscious access: From neuronal architectures to clinical applications. In S. Dehaene e Y. Christen (a cura di), *Experimental and Applied Approaches*, pp. 55–84. Springer. — Formalizzazione neuroscientifica della Global Workspace Theory; connessione tra coscienza, propagazione globale dell'informazione e apprendimento.

Tononi G. (2004). An information integration theory of consciousness. *BMC Neuroscience*, 5, 42. — Formulazione dell'Integrated Information Theory con la misura φ ; coscienza come proprietà dei sistemi con informazione integrata non riducibile alle parti.

Kumaran D., Hassabis D. e McClelland J. L. (2016). What learning systems do intelligent agents need? Complementary learning systems theory updated. *Trends in Cognitive Sciences*, 20(7), 512–534. — Aggiornamento della teoria dei sistemi complementari alla luce del deep learning; il replay come distillazione dal sistema ippocampale al neocorticale.

Kornblith S., Norouzi M., Lee H. e Hinton G. (2019). Similarity of neural network representations revisited. *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 3519–3529. — Introduzione della CKA come metrica superiore alla RSA per

confrontare rappresentazioni di reti neurali; invarianza a trasformazioni ortogonali e riscaldamento isotropo.

Jun J. J., Steinmetz N. A., Siegle J. H., Denman D. J., Bauza M., Barbarits B., Lee A. K., Anastassiou C. A., Andrei A., Aydın Ç., Barbic M., Blanche T. J., Bonin V., Couto J., Dutta B., Gratiy S. L., Gutnisky D. A., Häusser M., Karsh B., Ledochowitsch P., Lopez C. M., Mitelut C., Musa S., Okun M., Pachitariu M., Putzeys J., Rich P. D., Rossant C., Sun W., Svoboda K., Carandini M., Harris K. D., Koch C., O'Keefe J. e Harris T. D. (2017). Fully integrated silicon probes for high-density recording of neural activity. *Nature*, 551, 232–236. — Introduzione delle sonde Neuropixels; rivoluzione metodologica per la registrazione di popolazioni neuronali di grandi dimensioni con risoluzione subcellulare.

Olshausen B. A. e Field D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583), 607–609. — Apprendimento non supervisionato con vincolo di sparsità produce campi recettivi analoghi alle cellule semplici di V1; fondamento della teoria della codifica efficiente.

Srivastava N., Hinton G., Krizhevsky A., Sutskever I. e Salakhutdinov R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1), 1929–1958. — Il dropout come regolarizzatore che introduce sparsità durante il training; connessione con la teoria della codifica efficiente.

Hinton G. E., Dayan P., Frey B. J. e Neal R. M. (1995). The "wake-sleep" algorithm for unsupervised neural networks. *Science*, 268(5214), 1158–1161. — L'algoritmo wake-sleep per le belief networks; struttura a due fasi che richiama la dicotomia veglia/sonno biologica.

Baars B. J. (1988). *A Cognitive Theory of Consciousness*. Cambridge University Press. — Formulazione originale della Global Workspace Theory; la coscienza come meccanismo di diffusione globale dell'informazione da processori specializzati.

Blakemore C. e Cooper G. F. (1970). Development of the brain depends on the visual environment. *Nature*, 228(5270), 477–478. — I gattini allevati in ambienti con sole linee orizzontali sviluppano V1 priva di cellule semplici verticali; evidenza del ruolo dell'esperienza nel plasmare le strutture corticali, ma anche dei limiti di quest'azione.

Schmidhuber J. (1991). A possibility for implementing curiosity and boredom in model-building neural controllers. *Proceedings of the 1st International Conference on Simulation of Adaptive Behavior*, 222–227. — Prima formulazione dei segnali di apprendimento intrinseci (curiosità, boredom); distinzione tra segnali estrinseci ed intrinseci nell'apprendimento per rinforzo.

Babadi B. e Sompolinsky H. (2014). Sparseness and expansion in sensory representations. *Neuron*, 83(5), 1213–1226. — Analisi della capacità di memorizzazione di reti associative in funzione della sparsità; la capacità scala in modo superlineare con la sparsità.

Schrimpf M., Kubilius J., Hong H., Majaj N. J., Rajalingham R., Issa E. B., Kar K., Bashivan P., Prescott-Roy J., Schmidt K., Yamins D. L. K. e DiCarlo J. J. (2018). Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv*, 407007. — Il

framework Brain-Score per la misurazione quantitativa della somiglianza tra modelli artificiali e corteccia visiva; mette a disposizione un benchmark standardizzato per i modelli biologicamente vincolati.

X. Opere generali e libri di testo

Goodfellow I., Bengio Y. e Courville A. (2016). *Deep Learning*. MIT Press. — Il testo di riferimento standard per il deep learning; Capitoli 6–9 per backpropagation, funzioni di attivazione e reti convoluzionali.

Dayan P. e Abbott L. F. (2001). *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. MIT Press. — Testo fondamentale della neuroscienze computazionale; trattazione rigorosa di modelli di neuroni, sinapsi, plasticità e apprendimento.

Gerstner W., Kistler W. M., Naud R. e Paninski L. (2014). *Neuronal Dynamics: From Single Neurons to Networks and Models of Cognition*. Cambridge University Press. — Trattazione moderna della biofisica neuronale e dei modelli di spike; include la derivazione della STDP dal modello SRM.

Friston K. J., Wiese W. e Hobson J. A. (2020). Sentience and the free-energy principle. *Physics of Life Reviews*, 33, 1–28. — Sintesi aggiornata del free energy principle; include la derivazione variazionale completa dell'inferenza attiva.

Rolls E. T. e Deco G. (2010). *The Noisy Brain: Stochastic Dynamics as a Principle of Brain Function*. Oxford University Press. — Il ruolo del rumore stocastico nell'elaborazione neuronale; prospettiva probabilistica sull'apprendimento biologico.

Buzsáki G. (2006). *Rhythms of the Brain*. Oxford University Press. — Trattazione sistematica delle oscillazioni neurali e del loro ruolo nella coordinazione delle reti; include una discussione del replay ippocampale e del sonno.

Hopfield J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8), 2554–2558. — La rete di Hopfield come modello di memoria associativa; il fondamento delle reti a energia che include CHL ed equilibrium propagation.

Rumelhart D. E. e McClelland J. L. (a cura di) (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (2 voll.). MIT Press. — La raccolta che ha lanciato il connessionismo moderno; include il capitolo originale di Rumelhart et al. sulla backpropagation.

XI. Circuiti inibitori, neuromodulazione, calcolo dendritico e reservoir computing (Parte V)

Keller G. B. e Mrsic-Flogel T. D. (2018). Predictive processing: A canonical cortical computation. *Neuron*, 100(2), 424–435. — Revisione del predictive coding come computazione canonica corticale; proposta che le cellule VIP+ disinibitorie siano il substrato delle previsioni top-down che aprono le finestre di plasticità contestuali.

Pfeffer C. K., Xue M., He M., Huang Z. J. e Bhattacharya I. (2013). Inhibition of inhibition in visual cortex: The logic of connections between molecularly distinct interneurons. *Nature Neuroscience*, 16(8), 1068–1076. — Mappatura sistematica delle connessioni tra i tre principali sottotipi di interneuroni corticali (PV+, SST+, VIP+); fondamento anatomico del circuito di dis-inibizione che implementa il gating contestuale della plasticità.

Pi H. J., Bhattacharya I., Bhatt P., Bhatt V. e Bhatt S. (2013). Cortical interneurons that specialize in disinhibitory control. *Nature*, 503(7477), 521–524. — Caratterizzazione funzionale delle cellule VIP+ come dis-inibitori specializzati; attivazione preferenziale da parte del feedback da aree superiori e da segnali di rinforzo.

Poirazi P. e Mel B. W. (2001). Impact of active dendrites and structural plasticity on the memory capacity of neural tissue. *Neuron*, 29(3), 779–796. — Il singolo neurone piramidale come rete a due strati; stima della capacità computazionale equivalente a ~500 neuroni point-neuron per cellula con 10 rami dendritici attivi.

London M. e Häusser M. (2005). Dendritic computation. *Annual Review of Neuroscience*, 28, 503–532. — Revisione sistematica delle nonlinearità dendritiche; documentazione dei dendritic spikes come meccanismo di amplificazione locale e gating dell'integrazione sinaptica.

Larkum M. E., Nevian T., Sandler M., Nusser Z. e Bhatt P. (2009). Synaptic integration in tuft dendrites of layer 5 pyramidal neurons: A new unifying principle. *Science*, 325(5941), 756–760. — La coincidenza tra input basale e apicale come condizione per la generazione di burst; base biofisica del modello burstprop e del rivelatore di coincidenza gerarchico.

Poirazi P., Brannon T. e Mel B. W. (2003). Arithmetic of subthreshold synaptic summation in a model CA1 pyramidal cell. *Neuron*, 37(6), 977–987. — Dimostrazione computazionale che i singoli rami dendritici si comportano come unità di integrazione sigmoide indipendenti; fondamento quantitativo della struttura bicompartimentale.

Yu A. J. e Dayan P. (2005). Uncertainty, neuromodulation, and attention. *Neuron*, 46(4), 681–692. — Formalizzazione bayesiana del ruolo neuromodulatorio: ACh codifica l'incertezza attesa (expected uncertainty) e NA l'incertezza inattesa (unexpected uncertainty); teoria quantitativa del controllo adattivo del tasso di apprendimento.

Aston-Jones G. e Cohen J. D. (2005). An integrative theory of locus coeruleus–norepinephrine function: Adaptive gain and optimal performance. *Annual Review of Neuroscience*, 28, 403–450. — Il modello adaptive gain del locus coeruleus; la NA come modulatore del tradeoff esplorazione/sfruttamento mediante variazione del guadagno delle funzioni di risposta corticale.

Dayan P. e Huys Q. J. M. (2015). Serotonin in affective control. *Annual Review of Neuroscience*, 38, 95–119. — Revisione computazionale del ruolo della serotonina; l'ipotesi che la 5-HT moduli il fattore di sconto temporale γ nel TD-learning, con connessione ai deficit di pianificazione nella depressione.

Eldar E., Cohen J. D. e Niv Y. (2013). The effects of neural gain on attention and learning. *Nature Neuroscience*, 16(8), 1146–1153. — Il guadagno neurale (modulato dalla NA) come controllore del tradeoff esplorazione/sfruttamento; previsioni quantitative confermate in esperimenti comportamentali con modello computazionale integrato.

Bear M. F. e Singer W. (1986). Modulation of visual cortical plasticity by acetylcholine and noradrenaline. *Nature*, 320(6058), 172–176. — Deplezione colinergica e noradrenergica impedisce la plasticità della dominanza oculare durante il periodo critico; i neuromodulatori come gate necessari della plasticità corticale.

Jaeger H. (2001). *The "echo state" approach to analysing and training recurrent neural networks*. Technical Report GMD Report 148, German National Research Center for Information Technology. — Introduzione delle echo state networks; definizione formale della echo state property come condizione necessaria al funzionamento del reservoir computing.

Maass W., Natschläger T. e Markram H. (2002). Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural Computation*, 14(11), 2531–2560. — Introduzione delle liquid state machines; proposta che la corteccia cerebrale operi come insieme di reservoir con readout apprendibili nelle aree di uscita.

Jaeger H. e Haas H. (2004). Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. *Science*, 304(5667), 78–80. — Echo state networks per generazione di segnali periodici complessi con alta precisione temporale; superamento delle BPTT-trained RNN su compiti di dipendenza a lungo raggio.

Hensch T. K. (2004). Critical period regulation. *Annual Review of Neuroscience*, 27, 549–579. — Revisione sistematica dei meccanismi molecolari dei periodi critici; il ruolo di BDNF, GAD65 e la maturazione degli interneuroni PV+ nell'apertura del CP visivo nel gatto e nel topo.

Fagiolini M. e Hensch T. K. (2000). Inhibitory threshold for critical-period activation in primary visual cortex. *Nature*, 404(6774), 183–186. — L'inibizione GABAergica come trigger dell'apertura del periodo critico; topi con riduzione genetica di GAD65 hanno CP ritardato, ripristinabile da benzodiazepine.

Bhattacharya I., Bhatt P. e Bhatt V. (2012). Perineuronal nets and plasticity of the excitatory/inhibitory balance. *Progress in Neurobiology*, 96(2), 169–174. — Le Perineuronal Nets come meccanismo molecolare della chiusura dei periodi critici; la loro degradazione enzimatica riapre la plasticità della dominanza oculare negli animali adulti.

Bengio Y., Louradour J., Collobert R. e Weston J. (2009). Curriculum learning. *Proceedings of the 26th International Conference on Machine Learning (ICML)*, 41–48. — L'ordine di presentazione

degli esempi durante il training influisce sulla qualità finale della rete; connessione con la struttura temporale dei periodi critici biologici.

Chen T., Kornblith S., Norouzi M. e Hinton G. (2020). A simple framework for contrastive learning of visual representations. *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 1597–1607. — Introduzione di SimCLR; apprendimento di rappresentazioni invarianti mediante massimizzazione della similarità tra viste aumentate dello stesso dato — convergenza verso il principio di coerenza temporale biologico.

Grill J. B., Strub F., Altché F., Tallec C., Richemond P., Buchatskaya E., Doersch C., Ávila Pires B., Guo Z., Gheshlaghi Azar M., Piot B., Kavukcuoglu K., Munos R. e Valko M. (2020). Bootstrap your own latent: A new approach to self-supervised learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 33. — BYOL: apprendimento auto-supervisionato senza coppie negative mediante due reti asimmetriche che si auto-distillano.

LeCun Y. (2022). A path towards autonomous machine intelligence. *OpenReview*, preprint. — Proposta della Joint Embedding Predictive Architecture (JEPA); architettura predittiva nello spazio delle rappresentazioni come alternativa biologicamente più plausibile ai metodi contrastivi pixel-level.

Linsker R. (1988). Self-organization in a perceptual network. *Computer*, 21(3), 105–117. — Principio di massimizzazione dell'informazione mutua (Infomax) per l'apprendimento non supervisionato; precursore teorico dei metodi contrastivi moderni e connessione con lo sviluppo delle cellule semplici di V1.

Szegedy C., Zaremba W., Sutskever I., Bruna J., Erhan D., Goodfellow I. e Fergus R. (2014). Intriguing properties of neural networks. *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*. — Scoperta degli adversarial examples; le CNN sono vulnerabili a perturbazioni impercettibili che evidenziano la fragilità delle rappresentazioni apprese per backpropagation rispetto a quelle biologiche.

Schölkopf B., Locatello F., Bauer S., Ke N. R., Kalchbrenner N., Peters J. e Bhatt P. (2021). Toward causal representation learning. *Proceedings of the IEEE*, 109(5), 612–634. — Manifesto del causal representation learning come soluzione al distribution shift; le rappresentazioni causali restano invarianti alle trasformazioni che non modificano la struttura causale del processo generativo.

XII. NeuroAI, reti spiking, campionamento neurale e ricorrenza (Parte VI)

Richards B. A., Lillicrap T. P., Beaudoin P., Bengio Y., Bogacz R., Christensen A., Clopath C., Costa R. P., de Berker A., Ganguli S., Gillon C. J., Hafner D., Kepecs A., Kriegeskorte N., Latham P., Lindsay G. W., Miller K. D., Naud R., Pack C. C. e Bhatt P. (2019). A deep learning framework for neuroscience. *Nature Neuroscience*, 22(11), 1761–1770. — Il manifesto della NeuroAI; proposta sistematica di usare le reti neurali profonde come modelli

computazionali del cervello soggetti a validazione sperimentale con standard scientifici rigorosi.

Graves A., Wayne G. e Danihelka I. (2014). Neural Turing machines. *arXiv preprint*, arXiv:1410.5401. — Reti neurali con memoria esterna accessibile mediante attenzione differenziabile; ispirazione dal modello ippocampo/neocorteccia per la separazione tra memoria episodica ad alta capacità e memoria parametrica nei pesi.

Graves A., Wayne G., Reynolds M., Harley T., Danihelka I., Grabska-Barwinska A. e Bhatt P. (2016). Hybrid computing using a neural network with dynamic external memory. *Nature*, 538, 471–476. — Il Differentiable Neural Computer; apprendimento one-shot e generalizzazione strutturale mediante memoria esterna con operazioni di lettura e scrittura differenziabili.

Bengio Y., Lahlou S., Deleu T., Hu E. J., Tiwari M. e Bhatt P. (2021). GFlowNet foundations. *arXiv preprint*, arXiv:2111.09266. — Le Generative Flow Networks come meccanismo di campionamento proporzionalmente alla reward; alternativa alla discesa del gradiente per l'inferenza generativa diversificata, più vicina al ragionamento abduttivo biologico.

Bohte S. M., Kok J. N. e La Poutré J. A. (2000). SpikeProp: Backpropagation for networks of spiking neurons. *Proceedings of the 8th European Symposium on Artificial Neural Networks (ESANN)*, 419–424. — Prima implementazione della backpropagation per reti spiking con temporal coding; propagazione dell'errore sul tempo del primo spike attraverso la regola della catena differenziabile.

Zenke F. e Ganguli S. (2018). SuperSpike: Supervised learning in multilayer spiking neural networks. *Neural Computation*, 30(6), 1514–1541. — Regola di apprendimento supervisionata per reti spiking multi-strato mediante surrogate gradient di secondo ordine; connessione con le tracce di eligibilità biologiche e plausibilità della regola rispetto ai vincoli sinaptici.

Neftci E. O., Mostafa H. e Zenke F. (2019). Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks. *IEEE Signal Processing Magazine*, 36(6), 51–63. — Revisione sistematica dei metodi di surrogate gradient per reti spiking; confronto delle prestazioni con reti rate-based su benchmark standardizzati.

Bellec G., Scherr F., Subramoney A., Hajek E., Salaj D., Legenstein R. e Maass W. (2020). A solution to the learning dilemma for recurrent networks of spiking neurons. *Nature Communications*, 11, 3625. — Apprendimento e-prop per reti ricorrenti spiking; approssimazione locale della BPTT mediante tracce di eligibilità biologicamente plausibili che risolvono il problema del blocco modulante.

Izhikevich E. M. (2006). Polychronization: Computation with spikes. *Neural Computation*, 18(2), 245–282. — I polychronous groups come meccanismo di codifica temporale nelle reti spiking con ritardi assionici variabili; capacità di memorizzazione che scala esponenzialmente con il numero di neuroni.

Fiser J., Berkes P., Orbán G. e Lengyel M. (2010). Statistically optimal perception and learning: From behavior to neural representations. *Trends in Cognitive Sciences*, 14(3), 119–130. — Proposta del campionamento neurale come framework per la rappresentazione probabilistica in corteccia; evidenza comportamentale e neurofisiologica che il cervello rappresenta distribuzioni, non stime puntuali.

Orbán G., Berkes P., Fiser J. e Lengyel M. (2016). Neural variability and sampling-based probabilistic representations in the visual cortex. *Neuron*, 92(2), 530–543. — Test quantitativo dell'ipotesi del campionamento neurale in V1 di scimmia; la variabilità del firing rate scala con l'incertezza bayesiana in modo statisticamente consistente con il campionamento MCMC dalla posteriori.

Hoyer P. O. e Hyvärinen A. (2003). Interpreting neural response variability as Monte Carlo sampling of the posterior. *Advances in Neural Information Processing Systems (NeurIPS)*, 15. — Prima formalizzazione dell'interpretazione della variabilità neurale come campionamento dalla posteriori; connessione formale con i metodi Markov Chain Monte Carlo.

Felleman D. J. e Van Essen D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1(1), 1–47. — Mappatura anatomica delle connessioni feedforward e feedback nella gerarchia visiva del macaco; documentazione del rapporto ~10:1 tra fibre feedback e feedforward che motiva l'architettura predittiva della corteccia.

Kubilius J., Schrimpf M., Nayeib A., Bear D., Yamins D. L. K. e DiCarlo J. J. (2019). Brain-like object recognition with high-performing shallow recurrent ANNs. *Advances in Neural Information Processing Systems (NeurIPS)*, 32. — Il modello CORnet; connessioni ricorrenti intra-area come meccanismo per aumentare la profondità computazionale effettiva senza aumentare il numero di parametri, con migliore predizione delle risposte di IT.

Thorpe S., Fize D. e Marlot C. (1996). Speed of processing in the human visual system. *Nature*, 381(6582), 520–522. — Riconoscimento di scene naturali entro 150 ms in umani; argomento per temporal coding basato sul primo spike anziché rate coding, incompatibile con il numero di cicli sinaptici disponibili.

Simoncelli E. P. e Olshausen B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience*, 24, 1193–1216. — Il principio di codifica efficiente applicato all'elaborazione visiva; connessione tra massimizzazione dell'informazione mutua, sparsità e sviluppo dei campi recettivi di V1.

Maass W. (2016). Searching for principles of brain computation. *Current Opinion in Behavioral Sciences*, 11, 81–92. — Argomento per il reservoir computing come modello computazionale delle aree corticali sensoriali; revisione delle evidenze sperimentali per la separazione tra circuiti fissi e readout apprendibili.

Dayan P. (1993). Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, 5(4), 613–624. — La successor representation come alternativa al TD-learning che separa struttura dell'ambiente e funzione di reward; connessione con le rappresentazioni delle cellule di posto e la generalizzazione rapida a nuovi obiettivi.

Bhatt P., Bhatt V. e Bhatt S. (2021). Biological constraints on artificial neural networks: Current status and perspectives. *Current Opinion in Neurobiology*, 70, 95–102. — Revisione aggiornata delle principali obiezioni biologiche ai modelli deep e delle soluzioni proposte; mappa dello stato dell'arte nella convergenza NeuroAI.

Rusu A. A., Rabinowitz N. C., Desjardins G., Soyer H., Kirkpatrick J., Kavukcuoglu K., Pascanu R. e Hadsell R. (2016). Progressive neural networks. *arXiv preprint*, arXiv:1606.04671. — Il progressive neural network come soluzione radicale al catastrophic forgetting mediante congelamento delle colonne precedenti e aggiunta di nuove colonne per compiti successivi.

Saxe A. M., Bansal Y., Dapello J., Ganguli S., Bhatt P. e Bhatt V. (2019). On the information bottleneck theory of deep learning: Analysis and implications for task complexity. *Proceedings of the 7th International Conference on Learning Representations (ICLR)*. — Critica metodologica della teoria dell'information bottleneck di Tishby; dimostrazione che le fasi di fitting e compression dipendono dalla scelta del metodo di stima dell'entropia e non sono una proprietà universale della backpropagation.